Integrated Approaches to the Risk Prediction of First-episode Psychosis

Leirer, Daniel Jonathan

Awarding institution:
King's College London

Integrated Approaches to the Risk Prediction of First-episode Psychosis

Daniel Jonathan Leirer

Supervisor: Prof. Richard Dobson

Advisor: Prof. Robin Murray

Department of Biostatistics and Health Informatics
University of London, King's College London

This dissertation is submitted for the degree of
Doctor of Philosophy

King's College London March 2018

"Die Welt zu durchschauen, sie zu erklären, mag großer Denker Sache sein. Mir aber liegt einzig daran, die Welt und mich und alle Wesen mit Liebe und Wunder betrachten zu können."

Für Werner

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. The contents of this dissertation are the result of my own work, and are not the result of collaborations with others, unless specified in the text and acknowledgements in accordance with King's College London regulations. This PhD Thesis contains 30,000 words, excluding bibliography and appendix.

Daniel Jonathan Leirer
March 2018

Acknowledgements

This work would not have been possible without the contribution and support of many others. Foremost among them were my supervisors Richard Dobson, Steven Newhouse and Robin Murray, who allowed me to undertake this work and offered advice, guidance and support. I also would like to acknowledge Marta Di Forti, Diego Quattrone, Gerome Breen, Olesya Ajnakina and the entire GAP team without whom none of this would have been possible. I am especially grateful to Conrad Iyegbe who initially introduced me to this project, and offered his advice over several years. My deepest gratitude goes to Lorraine Gordon, Philip Asherson, Fruhling Rijsdijk and Baljinder Mankoo for their endless support when I needed it most. I would like to extend my gratitude to all the students and staff at the SGDP and IOPPN who have helped and kept me company over the last 4 years. Especially all the members of the Dobson lab.

A special thanks goes to all the friends, family and co-workers who provided encouraging words and commented on early drafts of this work. I would like to thank Werner Leirer, Michelle Leirer, Anika Oellrich and Laura Kutt for reading the first drafts and giving feedback. My thanks also go to Max Kerz for sharing ideas and providing motivation while writing. I would also like to thank my flatmates Marta Kutt, Neil Tan and Dickson Cheung for keeping my spirits up.

Finally, I would like to thank the Guy's and St'Thomas Charity and King's Bioscience Institute for generously supporting my studies and the MRC-Social, Genetic and Developmental Psychiatry Centre for its support.

Abstract

Psychosis is a complex condition that features in many psychiatric disorders, and significantly affects the quality of life for both patients and family members. As part of the Genetics and Psychosis (GAP) study, this thesis presents one of the largest blood gene expression datasets on first-episode psychosis patients to date. This work aimed to characterise the blood-based biological perturbations in psychosis and to investigate the predictive ability of gene expression data.

Firstly changes in expression, between healthy controls and first-episode psychosis patients was explored, to identify genes associated with psychosis. I identified hundreds of differentially expressed genes and found associations to pathways involved in transcription, oxidative stress and viral replication.

Secondly, network approaches were used to construct modules of genes based on co-expression. I identified modules correlated to the severity of positive symptoms, and enrichment for pathways associated with the stress response and multiple brain regions.

Thirdly regularised generalised linear models with bootstrapping were used to generate predictions based on combinations of gene expression, genetic and demographic data. The highest performance was found for models incorporating gene expression data, with minimal improvement using additional data. Prediction accuracy for identifying psychosis samples was found to increase with severity of positive symptoms in schizophrenia samples, but not in other psychoses.

Finally, machine learning methods were used on public schizophrenia gene expression data to build a variety of predictive models. These models were tested on the Genetics and Psychosis (GAP) gene expression data. The results show increased predictive performance on samples with a schizophrenia diagnosis, compared to other types of psychosis.

Overall the thesis presents work analysing a novel gene expression dataset. The results suggest that blood gene expression signatures are more predictive for positive symptoms in schizophrenia than for other psychoses. This work also highlights expression differences in innate immune pathways and the stress response.

Table of contents

List of figures

List of tables

List of Algorithms

List of Acronyms and Abbreviations

ADAR Adenosine Deaminase, RNA Specific

AF antipsychotic-free

AKT1 AKT Serine/Threonine Kinase 1

ANN Artificial Neural Network

AUC area under curve

BRC-MH Biomedical Research Centre for Mental Health

CAMP Cathelicidin Antimicrobial Peptide

CNR1 Cannabinoid-1 Receptor

CNV copy number variation

Dejong DeJong Chronic Schizophrenia cohort

DE differentially expressed

DGE differential gene expression

DSM Diagnostic and Statistical Manual of Mental Disorders

DZ dizygotic

FDR false discovery rate

FEP first-episode psychosis

FUCA1 Fucosidase, Alpha-L-1

GAP Genetics and Psychosis

GBM stochastic gradient boosting

GLMNET Lasso and Elastic-Net Regularized Generalized Linear Models

GRINA Glutamate Ionotropic Receptor NMDA Type Subunit Associated Protein 1

GWAS genome-wide association study

HC healthy controls

HINT1 Histidine Triad Nucleotide Binding Protein 1

HPA hypothalamic–pituitary–adrenal

HSV-1 herpes simplex virus, type 1

IOPPN Institute of Psychiatry, Psychology and Neuroscience

ICD-10 International Classification of Diseases 10

KNN k-Nearest Neighbours

LCL lymphoblastoid cell lines

LCN2 Lipocalin 2

MDD Major Depressive Disorder

MHC major histocompatibility complex

MRC Medical Research Council

MZ monozygotic

NMDAR *N*-methyl-D-aspartate receptor

NF-κB nuclear factor kappa-light-chain-enhancer of activated B cells

OOB out of box

OPCRIT Operational Criteria Checklist for Psychotic Illness and Affective Illness

PANSS Positive and Negative Syndrome Scale

PBMC peripheral blood mononuclear cells

PGC Psychiatric Genomics Consortium

PRKY Protein Kinase, Y-Linked, Pseudogene

PKU Phenylketonuria

PRR pattern recognition receptors

PRS polygenic risk score

RBCK1 RANBP2-Type And C3HC4-Type Zinc Finger Containing 1

RF random forest

RFE recursive feature elimination

ROC receiver operating characteristic

SGDP Social, Genetic & Developmental Psychiatry Centre

SLaM South London and Maudsley

SVA Surrogate Variable Analysis

SVM Support Vector Machines

SNP single nucleotide polymorphism

TFRC Transferrin receptor 1

WGCNA weighted gene co-expression network analysis

XIST X Inactive Specific Transcript

Chapter 1

Introduction

Psychosis is a severe psychiatric condition affecting approximately 4 in 1000 individuals (Jb et al., 2012) per year in the UK. The impact on quality of life, relationships and health services is significant. Successful diagnosis and treatment is often a long process with patients sometimes receiving multiple diagnoses and medications over several years before finding a combination that works if such a combination is found at all.

Over the recent decades, it has become clear that psychosis is caused by a complex interplay between genetics and the environment. Transcriptomic approaches have been used to capture this interaction, to understand psychosis at a molecular level and to find biomarkers. However, few studies have looked at first episode psychosis.

This thesis will present work on the largest, to our knowledge, gene expression cohort in first episode psychosis.

In this chapter, I will give a general overview of psychosis, before discussing the literature on heritability, genetics and environmental risk factors. I will then review the gene expression and classification literature on this topic. Finally, I will discuss the aims of this thesis, and provide an overview of the remaining chapters.

1.1 Psychosis

Psychosis in its broadest sense describes a condition which is defined by a loss or distortion of reality. In a medical setting, psychosis becomes pathological if it significantly interferes with the quality of life of individuals affected. Patients suffering from psychosis often experience symptoms that include hallucinations, delusions, paranoia and thought disorders, which can wreak havoc on their ability to function in everyday life. This in combination with the still considerable stigma associated with the condition, leads to a poor long-term outcome, which in turn places a considerable burden on relatives, healthcare providers and

the economy (Awad and Voruganti, 2008; O'Malley et al., 2011). In general, psychosis can occur in the context of severe mental illness (such as Schizophrenia and Bipolar Disorder), neurodegenerative diseases (such as Alzheimer's disease) and due to pharmacologically active compounds, such as amphetamines. This thesis focuses specifically on psychosis in mental illness and the psychiatric context.

Historically psychiatric thinking on psychosis has been heavily influenced by the German school of thought led by Emil Kraepelin who firmly believed in discreet classifications of psychiatric illness. He famously introduced the Kraepelinian system for classifying mental health disorders in 1898, in which he described what he thought were the two principal categories of psychosis, namely dementia praecox (later renamed to schizophrenia) and manic-depressive disorder (encompassing all affective disorders).

Since Kraeplin two main systems of classification have emerged which aim to classify mental illness. The Diagnostic and Statistical Manual of Mental Disorders (DSM), which is now in the 5th edition (American Psychiatric Association. and American Psychiatric Association. DSM-5 Task Force., 2013), is primarily used by services in the United States. The other system is the International Classification of Diseases 10 (ICD-10) compiled by the World Health Organisation (WHO) (Organization, 1992) which aims to contain classifications for all disorders including mental and behavioural ones.

Both the DSM-5 and ICD-10 have significantly expanded categories of mental disorders, which has been controversial and led to confusion in assessing the literature on specific conditions as the definitions have often changed dramatically over the last 100 years and are continuing to do so. Recent evidence suggesting an overlap between severe mental health problems such as schizophrenia and bipolar disorder (discussed later), has led to research that focuses on symptoms rather than diagnostic classification. This thesis, therefore, focuses primarily on non-organic first episode psychosis, in an attempt to bypass inconsistencies in diagnosis.

In this section, I will give a general overview of psychosis symptoms, the main psychiatric diagnoses that psychosis patients may receive and the rationale for focusing on first episode psychosis sufferers in this thesis.

1.1.1 Symptom Presentation in Psychosis

The definition used for psychosis in this thesis can be summed up as including the Schizophrenia, Schizotypal and Delusional Disorders (ICD-10 codes F20-F29) and Mood disorders (ICD-10 codes F30-F33). For Mood disorders, this encompasses Manic episodes, Bipolar disorder and Major depression. Symptom presentation can look radically different from patient to patient, leading to the original attempts to classify psychiatric conditions. Symptoms are

routinely categorised into Positive, Negative and Psychopathology scales as is done with the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987). Here positive symptoms, in a broad sense, refer to increases in thoughts and behaviours such as with hallucinations, delusions, excitement or grandiosity. In contrast, negative symptoms are characterised by a reduction in actions, emotions, social interactions, reactivity or thinking. Psychopathology symptoms can include a larger range of symptoms, that can include guilt, depression, poor judgement and impulse control or anxiety. None of these lists of symptoms is exhaustive, however.

Table 1.1 Symptom Presentation in Psychiatric Disorders

ICD-10	Diagnosis	Symptoms Presentation: Frequency			
Schizophrenia and delusional disorders		Hallucinations	Delusions	Catatonia	Thought Disorder
F20	Schizophrenia	Often	Often	Often	Often
F22	Delusional Disorder	Sometimes	Always	Never	Rarely
F25	Schizoaffective Disorder	Sometimes	Sometimes	Sometimes	Sometimes
F29	Unspecified Psychosis	Often	Often	Sometimes	Often
Mood (affective) disorders		Hallucinations	Delusions	Catatonia	Thought Disorder
F30	Manic Episode	Often	Often	Never	Often
F31	Bipolar Disorder	Sometimes	Sometimes	Rarely	Often
F32	Major Depression	Rarely	Rarely	Rarely	Sometimes

Table of diagnostic categories according to the ICD-10, with expected frequency of psychotic symptoms, based on diagnostic criteria. While affective disorders do feature psychotic symptoms, they are not as common.

Schizophrenia

Schizophrenia is the most frequent diagnosis in patients with psychosis symptoms included in this study and research on schizophrenia has led to some of the most significant results in psychiatric biology over the last decade (discussed later). Schizophrenia is one of the most disabling conditions discussed here, featuring prominent positive and negative symptoms in addition to a range of psychopathology ones. While recovery is possible many patients struggle their entire life with symptoms that can make it impossible to have lasting relationships or jobs. Schizophrenia onset peaks in the late twenties, but can start to develop in the late teens, and all throughout life. More men than women are on average affected, although women have a second peak for age of onset around menopause (Häfner, 2003).

Much of the literature discussed in this thesis focuses on Schizophrenia since mood disorders don't always feature psychotic symptoms, at least at the level of clinical significance.

Bipolar Disorder and Major Depression with psychotic features

Bipolar disorder is characterised by at least one episode of mania or hypomania, and usually depression.

Mania is an elevation in the mood which at the far end can resemble symptoms of stimulant use such as high energy levels, losing the need for sleep, severe loss of insight, racing thoughts, hallucinations and grandiose behaviour and beliefs. Elevated moods, are often characterised by high sociability and goal-oriented behaviour, but can also prominently feature irritability and anger.

Bipolar mania can be devastating, with patients impulsively acting in uncharacteristic ways, starting companies, gambling, or starting romantic relationships which for many results in financial hardship or significant problems in personal relationships. Even though medications such as mood stabilisers are in general more tolerable in their side effects, many patients do not comply with their prescriptions since mania in its early stages can be extremely pleasurable.

Depression in isolation or as part of bipolar disorder can also feature psychosis, most commonly in the form of hallucinations and unusual beliefs, this is less common than in mania and schizophrenia, but is thought to affect up to 20% of patients. Also, episodes of bipolar disorder can feature both manic and depressed elements, which is referred to as mixed episodes. Depression in one form or another is among the most common mental health problems affecting up to 30% of the population at one point or another. In severe cases, patients don't have the energy to work, socialise or take care of themselves. This can reinforce depression, and lead to self-harming behaviour or even suicide.

Treatment for depression in the form of antidepressants has modest effects, with meta-analyses (Kirsch, 2014; Kirsch et al., 2008) finding little evidence of widespread benefits over placebo treatment. It is unclear if this merely reflects a heterogeneity in the underlying causes of depression or a deeper problem with the medication. Another factor that has been suggested is that misdiagnosis of bipolar disorder sufferers might skew results since some evidence suggests that antidepressants could increase the risk of mania and suicide.

Other Psychoses

In addition to Schizophrenia, Bipolar disorder and Major Depression there are three other diagnoses received by patients in this thesis.

The most common diagnosis is schizoaffective disorder which features prominent symptoms found in both schizophrenia and affective disorders. In general schizoaffective disorder is divided into two main subcategories based on the type of affective symptoms. These are schizoaffective manic subtype and depressed subtype.

Schizophreniform is a disorder introduced by the DSM and refers to a schizophrenia subtype. While schizophreniform suffers often receive a diagnose of schizophrenia if the symptoms become chronic or more severe, the outcome is often better, and recovery happens more frequently.

Finally, Delusional disorder represents a psychotic disorder characterised, as the name suggests, by delusions. These can take a variety of forms, from grandiose to persecution, but patients can not meet the criteria for schizophrenia or mania, meaning that negative and affective symptoms are not prominent parts of the disorder.

1.1.2 First Episode Psychosis

Individuals are defined as having first-episode psychosis (FEP) when they first come in contact with health services for psychosis. This means that initial diagnosis is unclear, and appropriate treatment can be difficult, in part because psychosis can be drug-induced or a symptom of non-psychiatric health conditions. The main problem, however, is that response to treatment, even with a diagnosis, often involves numerous medication trials per patient.

Most studies in psychiatry are limited to studying patients who are in a remission or are stabilised by medication, sometimes over many years. FEP patients are drug naive at first contact, and while stabilisation with medication is often necessary to get informed consent for inclusion in a study the effects of chronic medication use are significantly reduced. As such, FEP cohorts provide the opportunity to study non-static variables affected by biology and environment, such as gene expression, or protein levels. Since antipsychotic medications can

substantially affect these biological signatures, studies of FEP patients can provide insight at a relatively early point when patients are minimally affected. This approach can help in identifying biomarkers and biological signatures that may be more specific to psychosis, than the results of studies at later stages.

From a scientific and clinical perspective, such a signature applied to FEP patients could improve early intervention and have a significant impact on the recovery, quality of life of patients as well as free up resources of medical services, by identifying patient subgroups from a gene-environment perspective.

1.2 Psychosis Risk Factors

Psychosis has historically been understood from a variety of perspectives. Early German psychiatrists such as Kraepelin strongly favoured biological underpinning. This fell out of favour in the mid-1900s in part due to disastrous psychiatric interventions which contributed to the rise of vocal anti-psychiatric sentiment. In the UK, R.D. Laing's views that psychosis can be understood in the context of victimisation that occurs in the nuclear family, became hugely influence (Laing et al., 2016), but this was ultimately rejected with the introduction of high-quality twin and genetic studies.

Today mental health is arguably best understood in the context of the biopsychosocial model of mental health, which tries to account for genetic, psychological and socio-cultural factors in the development of psychiatric diseases. In the case of the schizophrenia disease progression, such an integrated model is examined in Figure 1.1.

In this section, I will review the current evidence for biological and environmental risk factors for psychosis, primarily in the context of schizophrenia. I will then examine modern ways of synthesising these views in the light of current biopsychosocial perspectives.

Finally, I will consider the hypothalamic–pituitary–adrenal (HPA) axis in the context of psychosis, and the extent that the existing literature fits in with deregulation in this system.

1.2.1 Twin Studies and Heritability

Evidence for a genetic component in psychosis comes from three primary sources, namely family history, twin studies and genetic studies, especially genome-wide association study. The early evidence for a genetic component came from family studies, showing that mental health disorders run in families (Kendler, 1983) and family history is the most important risk factor for schizophrenia (Sullivan, 2005). This, however, is somewhat misleading, as most cases of schizophrenia occur in individuals who do not have a close family member who is

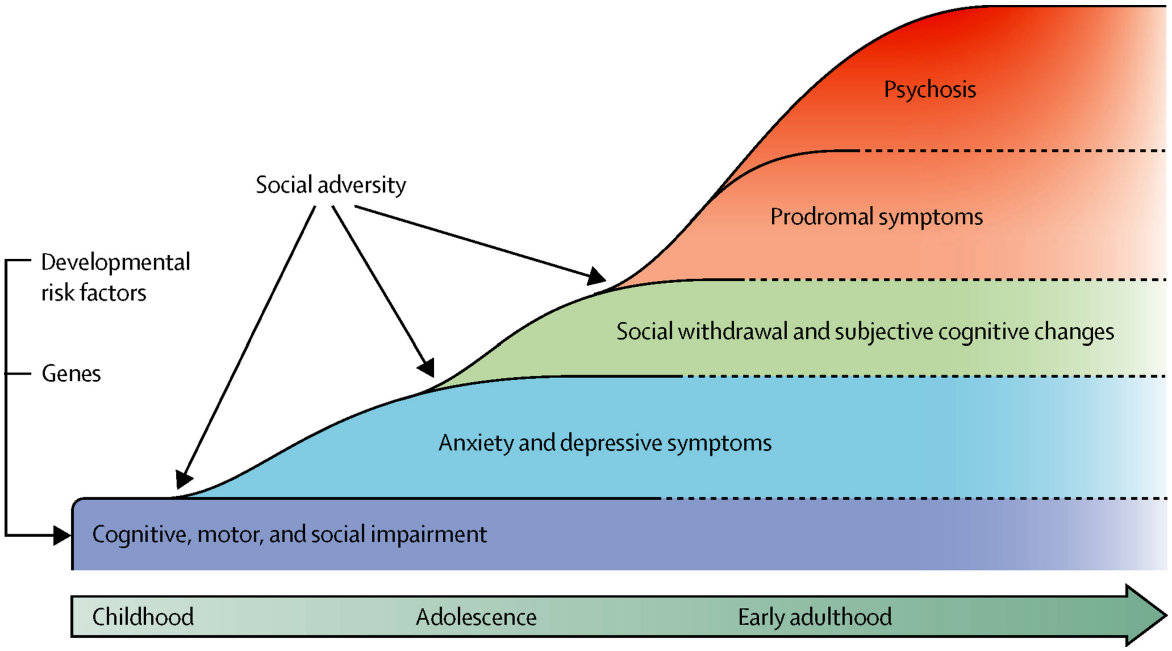


Figure 1.1 Schizophrenia Trajectory

Model of schizophrenia trajectory, showing progression of symptoms in the context of risk factors such as genetics, development and social adversity. *Reprinted from The Lancet, Volume 383, Howes, O. D. and Murray, R. M., Schizophrenia: an integrated sociodevelopmental-cognitive model, 1677-1678, Copyright 2014, with permission from Elsevier (Howes and Murray, 2014).*

affected. Family history alone is arguably just as compatible with a purely environmental model of psychosis, in which a disruptive family environment due to mentally ill relatives contributes to the development. In the remainder of this section, I will critically examine evidence from twin studies.

Twin Studies: Heritability findings for Psychiatric Conditions

Early twin studies significantly strengthened evidence of a genetic component for psychosis. The surprising results of these studies were the high heritability estimates for psychiatric conditions, especially in the case of bipolar disorder and schizophrenia. These studies suggested that genetics is the most significant factor and suggested little or no contribution of the shared environment (factors that would be shared by both twins). Instead, non-shared environment (environmental influences that are not shared by twins) surprisingly accounted for most and in the case of schizophrenia studies almost all environmental variation. One major study investigating schizophrenia found heritability estimate as high as 85% (Cardno et al., 1999), and a 2003 systematic meta-analysis, which pooled 14 previously published twin studies from 1941 to 1999 estimated heritability to be 81% (Sullivan et al., 2003). Estimates for Bipolar Disorder are similarly high with the largest twin studies finding heritability estimates between 60% and 87% (Smoller and Finn, 2003), while heritability for depression was found to be around 40% (Sullivan et al., 2000).

In the context of significant advances in DNA sequencing and microarray technology, and the human genome project nearing completion, these heritability estimates led to a focus on identifying the genetic and molecular underpinnings of schizophrenia and other psychiatric disorders. There are however some issues with this interpretation. In the next sections, I will review some of the problems and limitations of twin studies and heritability as a concept.

Twin Studies: Methodology

Using Twins to disentangle environmental and genetic traits was originally conceived by Francis Galton, and first implemented in the modern sense by Siemens (1924). While statistical approaches for analysis have become significantly more sophisticated, the underlying assumptions remain largely the same. Twin studies make use of the fact that monozygotic (MZ) and dizygotic (DZ) twins share 100% and 50% of their DNA respectively. In principle, all traits can be understood as resulting from a combination of genetics and environment. In twin studies this is usually represented with five variables; a for additive genetics, d for dominant genetics effects, c for shared environment e for non-shared environment and z for error in

measurement. To correctly account for all variables, twin studies require that twins are raised apart. This rarely happens, so most studies use the simpler ACE additive model, dropping *d*.

Twin studies using the ACE model make several core assumptions that are essential to their interpretation, as discussed by Rijsdijk and Sham (2002). Violation of these assumption calls into question the results for a given population. The most important of these, which will be discussed in more detail, is the assumption that MZ and DZ twins are equally affected by their shared environment insofar as it is relevant to the trait in question. DZ who have different genders will obviously not have the same extent of a shared environment which may contribute to psychological traits compared to MZ twins. While there are of course more significant reasons to use DZ twins that have the same gender, it illustrates an important point.

In the ACE model, the variation in phenotype is simply the sum of variation in additive genetics (A), shared environment (C) and non-shared environment (E). By calculating the intraclass correlation *r* between twins (which measures the concordance), and given that all assumptions are met (discussed later), the three variables are estimated by using Falconer's formula :

$$\textbf{Falconer's formula:} \tag{1.1}$$

$$\textit{Non-shared environment} : E = 1 - r(MZ) \tag{1.2}$$

$$\textit{Additive genetics} : A = 2(r(MZ) - r(DZ)) \tag{1.3}$$

$$\textit{Shared environment} : C = r(MZ) - A \tag{1.4}$$

The *r*(MZ) captures the combined effect of C and A, since discordance in MZ twins can only occur from the non-shared environment. Thus E is calculated by just subtracting 1 (since the model requires that variation does not exceed 100%) from the MZ correlation, with Huntington's in which concordance is 100% *r*(MZ) would be 1, resulting in E of 0. A is calculated by subtracting the DZ from the MZ correlation and multiplying the product by 2. Finally, C can be calculated by subtracting A from the MZ correlation.

Twin Studies: Criticism and Flaws

Heritability estimates based on the ACE model have been heavily criticised by Schönemann (1997) who argued that these models are constructed incorrectly and fail to correct for errors in measurements. Schonemann also argues that they are applied in a way that is mathematically invalid often leading to paradoxical results, like negative heritability or shared environment. This happens if *r*(MZ)-*r*(DZ) results in a value below 0 or above 0.5.

Consider a society in which MZ twins are required by law to get a small tattoo to distinguish them, while DZ can choose. Such a society may have an r(MZ) of 0.9 and an r(DZ) of 0.2 for tattoos. Using falconers formula we find that A = 1.4, C = -0.5 and E = 0.1. The result is that heritability of getting a tattoo is 140% which is clearly nonsense. This is now hidden within complex structural computer models, that prevent negative values, despite never adequately addressing the fundamental mathematical criticism.

When comparing the results of twin studies for a range of areas, it is striking that Twin studies seem to show that psychiatric and behavioural traits are excessively hereditary, as can be seen in Table 1.2. Schizophrenia here has the highest heritability out of all traits listed, and practically no contribution from the shared environment. Schizophrenia is on par with the "structure of the human eyeball" regarding genetic and environmental contributions.

In general shared environment seems to have little influence on psychiatric disorders, personality disorders and personality traits. This stands in contrast to many biological structures and diseases, such as Gout, Cystis Fibrosis, Diabetes, Endocrine Function, height, and Structures of the mouth and head, which attribute between 27% and 54% of trait variation to shared environmental factors. None of the psychiatric, or personality traits mentioned here have such high environmental contributions, even though these characteristics would reasonably be expected to be influenced significantly by the environment.

It should be noted that all of the results of twin studies simply provide a measure of variation in a phenotype of a particular population that is estimated to be due to the respective component of the ACE model. No trait or disease can be understood in isolation and is by necessity always the result of interactions of genetics with the environment. Even if in practice all environmental influences reasonably encountered, would lead to the same trait, as with Huntington's disease or eye colour, for example.

Heritability therefore as a measure of genetic contributions is usually meaningless without also considering the environmental context.

An extreme illustration of this is a monogenic disease like Phenylketonuria (PKU), which causes severe disability and early death, but is practically asymptomatic by eliminating phenylalanine from the diet. As such an environment with sufficient phenylalanine is required to have a disease. In a hypothetically isolated population homogeneous in PAH gene mutations (which is disrupted in PKU), and where phenylalanine is not found in the diet, it would be a purely environmental condition due to poisoning.

In the context of intelligence which also has very high heritability estimates, Turkheimer et al. (2003) showed that the heritability of IQ was close to zero in low-income families, while shared environment modulated 60% of the variation. The reverse was found in families of

Table 1.2 Results from Twin Study meta-analysis

Trait Studied	Heritability (A)	Environment (C)	*r*(MZ)	*r*(DZ)
Psychiatric Diseases				
Schizophrenia	0.77	0.013	0.7	0.3
Bipolar Affective Disorder	0.67	0.15	0.82	0.45
Mood Disorders	0.6	0.08	0.4	0.24
Recurrent Depressive Disorder	0.45	0.03	0.37	0.14
Depressive Episode	0.34	0.11	0.45	0.27
Personality Disorders				
Sleep Disorders	0.51	0.12	0.48	0.24
Conduct Disorder	0.49	0.18	0.66	0.43
Phobic Anxiety Disorders	0.49	0.15	0.57	0.29
Obsessive-Compulsive Disorder	0.45	0.15	0.53	0.3
Emotional Disorder (Childhood)	0.44	0.2	0.57	0.3
Eating Disorders	0.4	0.04	0.47	0.19
Personality				
Intelligence	0.67	0.12	0.69	0.45
Recreation and Leisure	0.54	0.18	0.59	0.3
Personality	0.44	0.13	0.47	0.23
Attitudes to Strangers	0.35	0.15	0.41	0.20
Societal Attitudes	0.3	0.2	0.45	0.29
Intimate Relationships	0.29	0.06	0.36	0.1
Biological Structure/Function				
Structure of Eyeball	0.73	0.02	0.73	0.37
Height	0.63	0.30	0.91	0.57
Structure of Brain	0.6	0.12	0.72	0.45
Menstruation Functions	0.6	0.04	0.63	0.25
Structure of Mouth	0.57	0.34	0.82	0.4
Structure of Head and Neck	0.44	0.27	0.8	0.44
Sleep Functions	0.34	0.13	0.51	0.26
Endocrine Function	0.32	0.34	0.54	0.38
Psychomotor Functions	0.31	0.16	0.43	0.23
Disease				
Development of Motor Function	0.74	0.10	0.83	0.49
Dementia in Alzheimers	0.63	0.12	0.86	0.5
Asthma	0.53	0.09	0.73	0.33
Type 1 Diabetes	0.5	0.54	0.76	0.43
Parkinson's Disease	0.36	0.01	0.51	0.34
Cystic Fibrosis	0.17	0.45	0.49	0.59
Gout	0.17	0.34	0.42	0.48

Table of meta analysis results from twin studies, in 5 categories. Traits are ordered form highest to lowest heritability for each category. The combined data from all studies is shown for heritability (A) and shared environment (C). Intraclass correlations for Monozygotic twins (*r*(MZ)) and Dizygotic twins (*r*(DZ)) are also present. Data taken from Polderman et al. (2015).

high socio-economic status. This understanding is essential for any interpretation of genetic influences in other personality and psychiatric disorders.

The further problem is that the equal environment assumption is violated for personality and psychiatric disorders (Fosse et al., 2015). That means it is evident that the shared environment of identical twins is more similar in ways that contribute to the development of psychiatric traits than it is for dizygotic twins. Environmental factors such as birth weight, infections, stress and importantly childhood adversity have been shown to contribute to the risk of psychosis (Iyegbe et al., 2014). It has been established that dizygotic twins are on average taller and heavier than monozygotic twins (Jelenkovic et al., 2015) even at birth. This can partially be attributed to the fact that dizygotic twins do not always share the same placenta, while identical twins do. Childhood adversity was specifically studied in the context of schizophrenia by Fosse et al. (2015), where they found a significantly higher rate of concordance for stressful life events, neglect, sexual abuse and bullying in identical twins when compared to dizygotic twins. It is easy to see why individuals that, for example, look identical would experience such adversities with a higher concordance rate.

In response to such criticism, some twin researchers argue that genetics determines how people look or act, and as such the social responses to individuals can be attributed to genetics. A twin study published in psychological medicine by Sartor et al. (2011) for example, claimed that the heritability of women being violently assaulted is over 40%, and argued that inherited personality traits likely modulate this. According to Sartor et al. (2011) these personality traits are what make people place themselves in situations that are more likely to lead to being assaulted.

Such a definition of heritability simply defines environmental influences out of existence and fails to provide any useful insights.

1.2.2 The Genetics of Psychosis

In this section, I will examine genetic studies, primarily in the context of schizophrenia. Early research in this area focused on candidate gene studies as well as copy number variation's (CNV). More recently genome-wide association studies (GWAS), which due to extensive worldwide collaborations have led to the development of novel bioinformatics methods, have had a major impact on our current understanding of the genetic architecture of schizophrenia. Finally, I will discuss the issue of missing heritability.

Candidate Genes and Copy Number Variations (CNV)

Genetic studies for psychiatric disorders have until recently been relatively unsuccessful. In the early 2000s, numerous candidate genes for schizophrenia were identified but failed to replicate consistently. With a large replication study finding no support for any of the 14 examined genes (Sanders et al., 2008).

Research on copy number variations has been more successful, with a large replication study finding support for 11 of 15 previously implicated CNV's for Schizophrenia (Rees et al., 2014). This is perhaps not surprising since duplications and deletions often span multiple genes, and have a much more significant impact than individual genes.

One of the most replicated CNV's implicated in schizophrenia is a deletion in 22q11.2, which usually spans between 30 to 40 genes (Kobrynski and Sullivan, 2007). This deletion is best know for causing DiGeorge syndrome, and is linked to a range of severe clinical symptoms which most commonly result in disruptions to normal development, cardiac function (Kobrynski and Sullivan, 2007) and immune function (Sullivan, 2004). In addition patients have a 25 fold increased risk of developing major psychiatric conditions (Shprintzen, 2008), with 10% (Schneider et al., 2014) to 31% of patients reported to suffer from schizophrenia. Notable among the deleted genes in that area is Catechol-O-methyl transferase (COMT) which plays a role in dopamine metabolism and has in some studies been linked to schizophrenia (Caspi et al., 2005) and bipolar disorder (Lelli-Chiesa et al., 2011). Although in the case of schizophrenia these results have not been consistently replicated (Zammit et al., 2007).

Genome Wide Association Studies

Initial results of genome-wide association study (GWAS) were not very encouraging and failed to identify significant genes associated with psychiatric disorders. It was not until Stefansson et al. (2009) brought together the major GWAS cohorts worldwide in 2009, that multiple regions achieved significance for the first time. This study had a total sample size of over 2000 subjects and over 13.000 controls. Despite that, a relatively small number of 7 genome-wide significant hits were identified, with the majority of these mapping to the major histocompatibility complex (MHC) region. Following this the Psychiatric Genomics Consortium (PGC) for schizophrenia has led the advance in this area and steadily increased sample sizes, culminating in a recent nature publication which included more than 36.000 cases and 113.000 controls (Ripke et al., 2014), and these figures are continually expanding. This study identified a total 108 genome-wide significant hits, although these results still only account for a fraction of the estimated heritability for schizophrenia.

Despite this, the GWAS results have led to some interesting findings regarding the genetic architecture of major psychiatric disorders. One landmark study used the polygenic risk score (PRS) of 5 disorders, including Bipolar Disorder, Major Depressive Disorder (MDD) and Schizophrenia and found that the PRS for these disorders had significant predictive power for the other disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). The highest predictive power was between Schizophrenia and Bipolar Disorder, while the predictive power of MDD (perhaps due to lower estimated heritability and greater heterogeneity) was less predictive, although still significant.

Missing Heritability

The problem of so-called missing heritability in GWAS studies has been the subject of much debate. One argument has been that a substantial amount of this missing variance, may be "hiding" in the already existing GWAS cohorts (Maher, 2008). This is because a GWAS analyses millions of single nucleotide polymorphisms and the multiple testing burden becomes extremely high, in these conditions. The assumption is that if schizophrenia is highly polygenic, the contribution of any one single nucleotide polymorphism (SNP) will be low, and the detection threshold given the high number of SNPs is too high for most genuine signals to be identified. Attempts to circumvent this constraint led to the discovery that by aggregating the effects of weakly associated variants into a PRS, a significant detection power is recoverable (Purcell et al., 2009a). Using PRS the recent PGC cohort has been able to account for at least 7% of the variation on the liability scale (Ripke et al., 2014). Unfortunately, PRS have little predictive value in individuals of African descent. This is in large part due to much greater genetic diversity (Lu et al., 2014) in populations of African origin.

In any case, the heritability estimates from these studies fall short of the 60%-85% values found by twin studies, and alternative methods such as GCTA come up with significantly smaller heritability estimates generally between 27% (Loh et al., 2015) and 20% (Gusev et al., 2014) for schizophrenia.

Even these heritability estimates face a problem, however. As Lander and Schork (1994) explained, genetic associations can arise from population stratification as well as genuine links between genotype and phenotype or linkage disequilibrium between the assayed and functional variant. To illustrate the potential problem of population stratification, the hypothetical example of studying chopstick proficiency has been used (Lander and Schork, 1994; Vilhjálmsson and Nordborg, 2012). This argument proposes that a GWAS on chopstick use in San Francisco would find a high heritability and a strong association with HLA-A1 (common in east Asians). In this case, chopstick use is merely correlated with genetic

markers, and heritability is measuring the environment being passed on, in the form of cultural transmission and self-identification. This problem of population stratification is well known and various methods have been developed to address this, such as EIGENSTRAT (Price et al., 2006). In fact both Lander and Vilhjalmsson, have argued that population stratification, while complex, can be adequately addressed. Nonetheless, this is an issue that has to be kept in mind when examining genetic data in the context of heritability, especially for complex traits.

Non of this invalidates the significance of genetic results, but simply underlines the importance of cautiously evaluating results. An example for this is seen in the results of largest GWAS for schizophrenia (Ripke et al., 2014), which found the most significant signals in the MHC class II region, which is the same genomic region that would cause problems in the above example.

That being said efforts have been made by Ripke et al. (2014) to account for stratification. In addition recent experiments have shown a plausible mechanism for how complement component 4 (C4), which is located in the MHC locus, could be linked to schizophrenia (Sekar et al., 2016). Over-expression of C4 led to increased synapse pruning in mice, during early development, which the authors suggest could explain structural brain differences found in some schizophrenia suffers, as well as cognitive deficits. Besides CNVs and major deletions, as in DiGeorge syndrome, could provide useful information about disrupted pathways at least for subsets of patients. Although Fosse et al. (2016) have argued, that the increased risk of schizophrenia and other psychiatric conditions, found in association with conditions such as DiGeorge syndrome might be better explained by significant disability and the resulting social adversity.

1.2.3 Environmental Risk Factors

There are a large variety of environmental risk factors for schizophrenia which can include neurodevelopmental factors, famine, childhood trauma, drug use, such as the smoking of cannabis, and even gender and socio-economic status (Iyegbe et al., 2014) (see Table 1.3 for a more complete list). One of the differences between environmental and genetic variables to risk is their effect size. Individual common genetic variants found in highly polygenic diseases tend to have at best a modest effect when expressed as odds ratios which are typically between 1.1 and 1.4. Copy number variations tend to have significantly higher effects, with a recent study estimating the odds ratios for CNVs to range from 2 to over 50 (Rees et al., 2014). Environmental risk factors, in contrast typically range between 1.5 and 11.0.

In addition lifetime prevalence is known to vary dramatically across gender, country and study, from 0.7 per 1000 to 12.8 per 1000 (McGrath et al., 2008), underlining the importance

of environmental context. This point is illustrated by one review showing that incidence of schizophrenia is significantly higher for migrants and their children, with Black and Pakistani individuals having 10 and 16 fold higher incidence rates of schizophrenia respectively(Jb et al., 2012).

One study examining environmental risk relating to ethnic minority status and urban environment, Kirkbride et al. (2010) argued that targeting these factors successfully in black and minority communities in England could prevent up to 22% of all psychotic illness and 27% in communities that are significantly affected. While this study was criticized for not controlling for family history (Iyegbe et al., 2014), it also did not include a series of other environmental factors that are only partially captured by ethnicity and urban status.

Given the much higher incidence of schizophrenia in migrants (Jb et al., 2012), and concordance rates for twins with schizophrenia falling well below 50% (Narayan et al., 2015), it is evident that the environment plays the dominant role in the development of psychiatric disorders, and needs to be carefully considered.

Reopening the nature vs nurture debate is not necessary for this, as it is apparent that every trait requires both. However genetic approaches have so far provided very little to help patients, and heritability estimates that encompass populations who experience radically different environmental pressures provide minimal useful information.

1.2.4 Gene Environment interactions

Turkheimer (2016) in a recent article while arguing that all traits are genetically influenced, expressed doubt that genetics can provide a coherent insight into sufficiently complex diseases or traits, by targeting problems at the wrong level of causality. He suggests that this will undoubtedly remain true for behaviours in society while expressing reservations for the usefulness in psychiatric disorders.

Despite this, approaches taking into account the environment and genetics have led to valuable insights that could have direct benefits for patients and inform policy. This is illustrated by the genetic interaction between cannabis use and genetics, where an AKT Serine/Threonine Kinase 1 (AKT1) variant was strongly implicated with a sevenfold increased risk of psychosis in cannabis users (Di Forti et al., 2012). This approach can, therefore, lead to testing of at-risk populations, and more targeted advice for patients.

In a similar manner finding biomarkers that can predict response to medications, such as antipsychotics or mood stabilisers, could substantially reduce exposure to the number of unsuccessful treatments.

Context	Environmental risk factor
Social	Urban-rural dwelling
	Social context – neighbourhood effects
	Social discrimination
	Migration
Environmental	Cannabis smoking
	Chemical pathogens
	Famine
Familial	Childhood trauma
	Childhood adversity
	Advanced paternal age
Neurodevelopmental	Seasonal birth
	Birth defects / obstetric complications
	Vitamin D
	Prenatal maternal infections
Economic	Developed vs developing country
	Socio-economic status
Other	Gender

Table 1.3 Environmental risk factors

Environmental risk factors for schizophrenia. Adapted from Iyegbe et al. (2014).

1.2.5 Stress Response and the HPA-axis

The stress response, since it can account for genetic and environmental factors, is one way to contextualise psychiatric disorders. A consistent finding across studies in the psychosis, and the wider psychiatric literature is immune deregulation. While the stress response, the immune system and disorders of the brain, such as psychosis seem fundamentally unrelated, this is not the case.

Chronic stress negatively impacts sleep which by itself can cause psychosis-like symptoms. Urbanicity, where sleep, especially in poorer neighbourhoods, may be affected, could potentially be mediated this way. Chronic stress can also cause damage to brain regions, in severe cases. Stress experienced by the mother during pregnancy, for example, has been shown to affect multiple brain regions including the amygdala, frontal cortex, and hippocampus (Weinstock, 2008) which are involved in emotion regulation, planning, and memory respectively. Chronic stress in adult anxiety disorder has also been suggested to contribute to damage in the hippocampus and frontal cortex (Mah et al., 2016). A proposed mechanism for this can be seen in Figure 1.2.

The immune system has been consistently implicated in schizophrenia studies, using omics data examining genetics (Ripke et al., 2014) as well transcriptomics (Gardiner et al.,

2013; Hess et al., 2016) and proteomics (Levin et al., 2010) studies. The transcriptomic and proteomic studies have found consistent up-regulation of the innate immune system. Such an upregulation can be modulated by genetic, environmental and psychosocial stressors. Viral and parasite infections, as well as toxins, have been linked to schizophrenia risk (Iyegbe et al., 2014), and they can upregulate the innate immune system via pattern recognition receptors (PRR) as seen in Figure 1.2. Psychosocial stressors can also have this effect, by activating the hypothalamic–pituitary–adrenal (HPA)-axis, where cortisol can be released in response to psychological factors.

Activation of the stress response via PRRs or the HPA-axis affects the quality of sleep, increases cortisol, down-regulates the adaptive immune system and up-regulates the innate immune system via TNF and IL-6. Since schizophrenia and mental health disorders, in general, occur more frequently in individuals with childhood adversity, trauma, certain infections and other external stressors this provides a plausible mechanism that can be mediated by genetics.

The HPA-axis plays a pivotal role in this process and has been a topic of attention for its role in psychosis. The HPA-axis fundamentally regulates cortisol production and down-regulation through a feedback loop. As the name suggests, it starts in the hypothalamus with corticotropin-releasing hormone (CRH) and vasopressin, which in turn stimulate the secretion of adrenocorticotropic hormone (ACTH) in the pituitary gland. This leads to glucocorticoid production in the adrenal cortex (cortisol among them), and a negative feedback loop in the pituitary and hypothalamus. Studies have also shown reduced hippocampal size in schizophrenia subjects, while this could be attributed to a variety of causes it is notable that the hippocampus has high levels of glucocorticoid receptors and individuals suffering from chronic stress show atrophy of the hippocampus.

There is debate over sex differences in cortisol production (Kirschbaum et al., 1992), with studies, in general, finding higher baseline cortisol levels in men, in addition to higher cortisol in response to acute stress, while women showed higher cortisol levels in the weeks following a chronic stressor (Paris et al., 2010). This may help explain sex differences in neuropsychiatric disorders between genders.

A recent systematic review by Karanikas et al. (2014), concluded that high blood cortisol levels in patients with first-episode psychosis (FEP) are a robust finding, while results for saliva cortisol levels were inconclusive. The authors suggested this may be due to differences in medication exposure in the latter case. In addition, longitudinal studies found that cortisol levels decreased after antipsychotic treatment. Karanikas et al. (2014) conclude that cortisol levels are up-regulated in psychotic episodes. The exact role of cortisol in psychiatric disorders is unclear, as results correlating cortisol levels with PANSS scores were

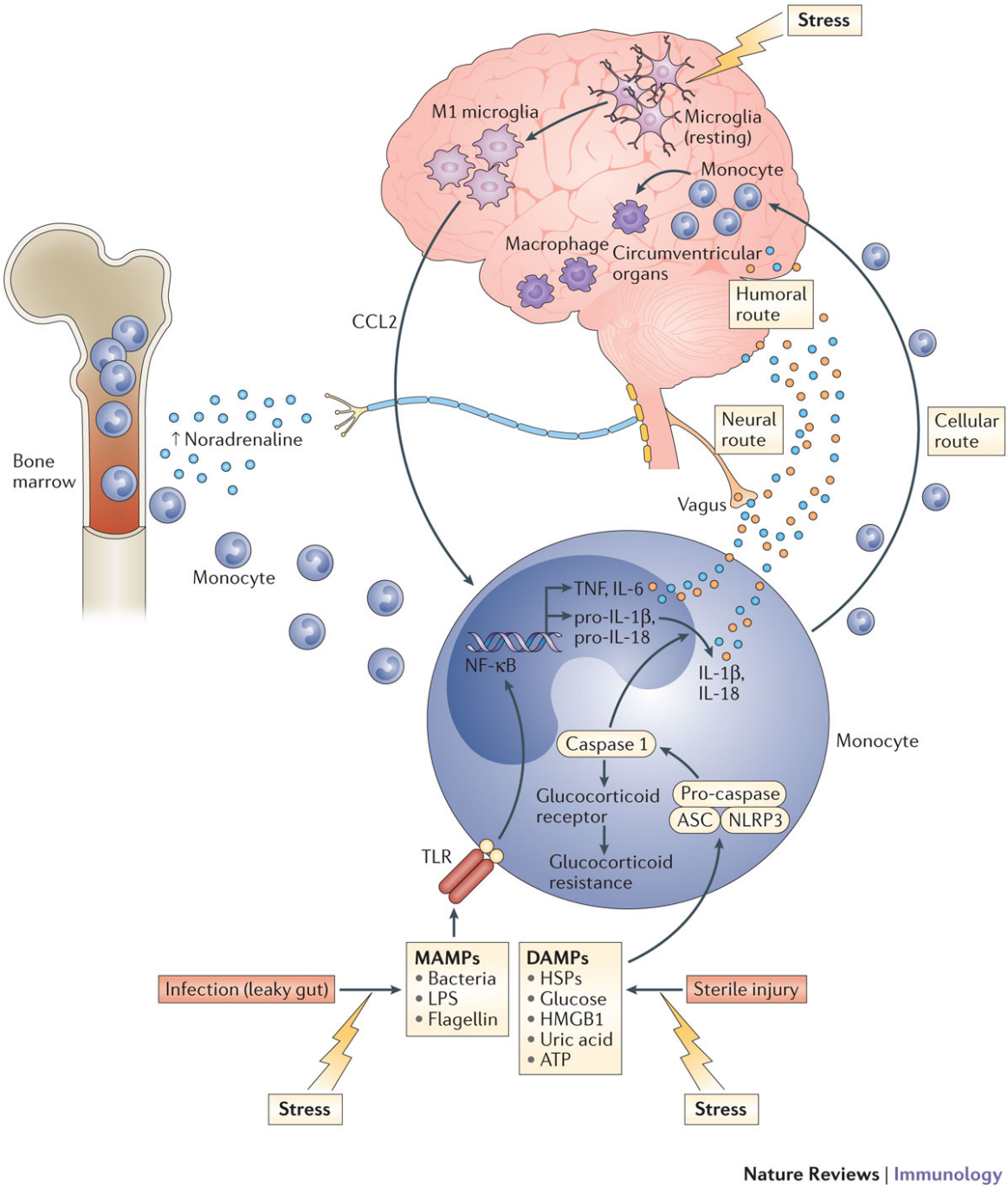


Figure 1.2 The Stress response and impact on the Brain

Figure detailing mechanisms by which stress can be transmitted to the brain causing damage. Psychosocial stressors activate the HPA-axis, leading to signal transduction via hormonal transmission to an upregulation of the innate immune system. This is also facilitated via infections or injury, by activation of pattern recognition receptors (PRR) such as Toll-like-Receptors (TLR). This leads to the activation of transcription factor NF-κB and TNF, IL-6 and IL-1 up-regulation. Activation of microglia in the brain, and macrophage migration can lead to inflammation and cell death in the brain, in chronic stress exposure. *Adapted by permission from Springer Nature: Nature Publishing Group, Nature Reviews Immunology, The role of inflammation in depression: from evolutionary imperative to modern treatment target, Miller, A.H. Raison, C.L., Copyright 2014, (Miller and Raison, 2016).*

contradictory, which may indicate that while stress and cortisol levels are high during an episode, they could be an acute response to the experience of symptoms.

Overall, the stress response and HPA-axis provide a framework which can account for a significant amount of the schizophrenia and psychosis literature. This also fits into a liability threshold model, based on biological, psychological, social and cultural factors, were a high genetic risk can trigger chronic stress and anxiety, leading to social withdrawal, sleep disturbances, and in severe cases damage to the brain, due to inflammation, oxidative stress and apoptosis.

1.3 Gene Expression Studies

While genetic information is stored in the form of DNA, most biological functions are performed by varying levels of enzymatically active proteins and to a lesser extent RNA (ribozymes). This is encapsulated by the central dogma of biology, introduced by Watson and Crick, which states that DNA stores information, transcribes genes to RNA, which ultimately gets translated into Proteins. To capture the interplay between genetics and the environment, gene expression or protein levels can be investigated. Both of these are non-static, and changes are the direct result of external stimuli, in contrast to DNA, which remains static (at least to the extent that needs to be considered in this thesis). This thesis will focus primarily on the transcriptome (global gene expression) in whole blood.

The literature on gene expression for schizophrenia and psychosis in general, while not small, is not comparable to the sample sizes found in genetics. Most expression studies have until recently, managed to recruit a few dozen patients at best. While more recent studies, and meta-analyses, have managed to increase those numbers, none have exceeded 500 patients. In comparison, GWAS studies have managed to secure genetic data from tens of thousands of patients. To further complicate this, differences in patient recruitment and methodology lead to significant problems in comparing and replicating results. Gene expression is significantly influenced by medication, environment, time of data collection, the microarray manufacturer and data processing to name just a few. While this provides many advantages, it also complicates interpretation and replication of results.

In this section, I will review the Gene Expression literature primarily in the context of schizophrenia. I will then examine the rationale of blood-based gene expression studies.

1.3.1 Gene Expression Studies of Schizophrenia and other Psychosis

A recent meta-analysis of transcriptomics in Schizophrenia identified 25 studies using blood (Hess et al., 2016) between 2005 and 2016. As noted before many findings in these studies have not been successfully replicated. Horváth and Mirnics (2015) have argued that even if issues with sample size, methodology and confounding variables are addressed, findings are likely to stay inconsistent due to Schizophrenia being a spectrum disorder. As such focus on differentially expressed genes is unlikely to provide much insight. They argue that understanding at the pathway level is required.

Studies using gene enrichment or network analysis have indeed led to some insights that replicate across studies, and some common themes have emerged. One study using peripheral blood mononuclear cells (PBMC) found upregulation of pathways in the innate immune system and RNA processing (Gardiner et al., 2013). This mirrors results from in blood-derived lymphoblastoid cell lines (LCL) from over 400 patients (Sanders et al., 2013) which also reported enrichment for immune-related function and miRNA processing. One study that used network analysis and was able to control for medication in chronic schizophrenia found that the most significant pathway was also related to the immune system, and that this pathway was enriched for brain-expressed genes (de Jong et al., 2012).

The most extensive gene expression study to date was conducted by Hess et al. (2016), and used data from 18 previously published studies (nine brain-based and nine blood-based). Weighted network analysis revealed significant upregulation of modules in blood associated with innate immune function (specifically TNF-alpha, NF-KB, P38 MAPK, IL-6, STAT3, LCN2, DEFA 1-4), cellular stress responses (hypoxia, UV exposure, unfolded protein response, apoptosis/p53 cascades), androgen response, RNA metabolism and oncogenesis. Downregulation of genes was prominent in genes located on chromosome 22q11.

These perturbations are not unique to schizophrenia, however, with expression studies in depression (Jansen et al., 2016), and PTSD (Breen et al., 2017) finding similar up-regulation in the innate immune system and specifically related to interferon I pathway. Something that is also supported by multiple brain-based studies in the case of schizophrenia (Arion et al., 2007; Hess et al., 2016; Mistry et al., 2013; Saetre et al., 2007).

Overall gene expression studies find consistent changes in pathways associated with the innate immune response, the stress response, mitochondrial dysfunction and miRNA regulation, across multiple conditions and diseases. Horváth and Mirnics (2015) suggest that environmental factors influence the same pathways for multiple psychiatric disorders and that genetic predisposition determines the disease. While this is plausible, they ignore the equally likely possibility that disease could be specified by the environment.

1.3.2 The rationale of using Blood for Transcriptomic studies in Psychiatry

Using blood-based studies for the psychiatric disorder may seem counter-intuitive, since these disorders are often presumed to be based in the brain. However, changes in the brain have detectable influences in periphery tissues. The best example is the fight-or-flight response, which starts with the perception of danger in the brain and is then translated into physiological responses, such as the secretion of adrenocorticotropic hormone into the bloodstream and is mediated by the HPA-axis or other stress responses as discussed earlier.

It is also important that biomarkers are readily accessible if they are to be clinically useful. While brain biopsies may be more likely to result in pathologically relevant markers, the invasiveness of such a procedure prohibits its use in a clinical setting. Additionally, blood-based studies have the advantage that patients can be followed up over an extended period. Thus non-specific blood-based biomarkers, which may not be of diagnostic value by themselves, could act as a starting point to justify more expensive and or invasive screening methodologies.

1.4 Machine Learning for Psychosis

One major issue in the treatment of psychosis-related disorders is that diagnosis is still based on somewhat outdated ideas regarding diagnosis groups. While there are practical reasons to use these classifications, incorporating biological markers as is done in the case *N*-methyl-D-aspartate receptor (NMDAR) antibody Encephalitis (Zandi et al., 2011) or DiGeorge syndrome for example. While these cases are easy to identify with biological markers, due to the straightforward causality, they represent important steps towards a more individualised and targeted treatment approach.

In the case of first episode psychosis were multiple pathways are likely disturbed by numerous unrelated mechanisms, a more sophisticated approach will be required. One possibility is using machine learning techniques, which have in recent years led to significant changes in numerous industries. Machine learning algorithms can "learn" from complex data and identify previously unknown groups or relationships by unsupervised learning, or build complex models to classify an input based on previously seen data.

A few attempts have been made to create classifiers for psychosis from transcriptomic data. One study using 26 first episode psychosis patients and controls matched by age, gender and ethnicity identified a 400 gene signature that can accurately distinguish patients and controls (Lee et al., 2012). They claimed a sensitivity and specificity of 100% and 96%

respectively. However, this study was severely limited since it did not replicate its findings, or even split their data into a training and testing group. Taking into consideration the use of 400 genes used to classify just 26 patients, it seems likely that they were simply overfitting data and these results are not applicable in other settings.

A more thorough approach to building a schizophrenia classifier by using blood-based gene expression data was performed by Takahashi et al. (2010). They used artificial neural networks in 52 antipsychotic-free schizophrenia patients and 49 controls. They also separated training and test data, which other studies (Lee et al., 2012; Middleton et al., 2005) failed to do. Using this approach, they achieved 87.0% accuracy in the test data. Nonetheless, these results need to be interpreted cautiously, something the authors freely admit, since sample size is still a significant issue.

A recent study by (Perkins et al., 2015) followed people with high-risk symptoms and built a classifier, to predict which individuals would develop psychosis. Their algorithm achieved an area under curve (AUC) of 0.88 and 0.83. Interestingly they achieved these results with just 15 and 6 transcripts, which included Lipoproteins, IL-1 Beta, Matrix metalloprotein 7, Apolipoprotein D, Factor VII, IL-7 and IL-8. Many of which have been linked to innate immune function, and neuropsychiatric disorders.

Hess et al. (2016) used nine publicly available gene expression datasets to build classifiers, resulting in the largest dataset for schizophrenia trained in this way. Since the available data came from two distinct microarray platforms (Illumina and Affymetrix), they trained their models on Illumina data and used Affymetrix datasets for validation. Using this approach, they achieved a predicted AUC between 90% and 99% in the training data and retained an AUC between 72% and 77% in the unseen test data. While all patients included had chronic schizophrenia and most were on antipsychotics, the robustness of these results is very encouraging.

While models would ideally also integrate gene expression, environmental and clinical variables, there is simply not enough data with this breadth at the moment. This does, however, provide an opportunity for further integration which will likely result in clearer signals and more relevant classifiers. The development of these classifiers has enormous potential benefits for the prevention or early treatment of people at risk of developing a psychotic illness. For this to become viable, the results of small exploratory studies need to be replicated and validated.

1.5 Aims

Since much of the literature on Psychosis focuses on Schizophrenia or Chronic Psychosis, and due to accumulating evidence of overlapping changes in core pathways shared by neuropsychiatric disorders, the primary aim of this thesis is to investigate if these patterns can be found in a first episode expression cohort. With the second arm of this thesis focusing on Machine Learning approaches.

Chapter 2 discusses methods and datasets used in this thesis. The GAP study is introduced, and both internal and external data is described, which includes transcriptomic, clinical, demographic and genetic information. The preprocessing pipeline, which was designed for gene expression, and the machine learning methods used in this thesis are covered in detail.

Chapter 3 describes the differential expression experiments between psychosis patients and healthy controls. Network analysis is also performed using weighted gene co-expression network analysis (WGCNA), to identify modules related to psychosis and symptom severity. Gene enrichment is used to identify enriched pathways in differentially expressed genes and modules.

Chapter 4 uses bootstrapping and machine learning to generate a series of classification models and classification frequencies for samples, based on combinations of gene expression data, PRS and demographic variables. Classification accuracy is assessed and compared to symptom severity and later diagnosis.

Chapter 5 details a machine learning based classification approach using an external gene expression dataset, first published by de Jong et al. (2012). The data is composed of samples taken from chronic schizophrenia patients and controls. Multiple machine learning algorithms are trained and merged into an ensemble model, to achieve higher classification accuracy. Accuracy is estimated in a validation dataset, and in the internal GAP expression data. Performance is compared in subgroups of the GAP data based on demographics, diagnosis and symptom severity.

Chapter 6 is the final chapter of this thesis providing an overall discussion of the results and their limitations.

Chapter 2

Methods

2.1 Datasets

The data used for the analysis presented in the following chapters is taken primarily from the GAP study conducted at the Institute of Psychiatry, Psychology and Neuroscience. This is supplemented by a publicly available gene expression cohort for replication purposes in chapter 5. In this section, I will give an introduction of these cohorts and provide an overview of the available data.

2.1.1 Genetics and Psychosis (GAP) study

The Genetics and Psychosis GAP study is the primary source of data for this thesis. The GAP study is an ongoing study, that is continuously recruiting patients, collecting follow-up data, and integrating additional data since 2005. Multiple teams across hospitals, institutes and disciplines are involved and have contributed to this study for over a decade. As such it is a highly valuable resource, which contains a large number of participants, and data on clinical, environmental and biological attributes. This section aims to provide information on 1.) the overall GAP cohort, 2.) Ethical approval and recruitment and 3.) a detailed overview of the Expression, Genetic and Patient data used in this thesis.

The GAP cohort consisted of 737 first-episode psychosis patients, 389 healthy controls and 48 mothers of patients, with varying levels of biological and environmental data.

Regarding biological data, blood and saliva samples have been collected for individuals, to generate "omics" data. This includes genetics, transcriptomics, proteomics and methylation.

The methylation and proteomics were not used in this thesis. The primary focus of this thesis is on transcriptomics from RNA microarrays (Illumina Expression array HT12-v4). Genome-wide SNP data (Illumina Human Exome Chip), processed by Dr Evangelos Vassos

(GAP Clinician), is also used. All transcriptomic data is derived from whole blood samples. Where possible genetic data comes from the same blood samples used for transcriptomics, otherwise saliva was used. Genetic and transcriptomic data was matched to individuals in collaboration with Dr Evangelos Vassos.

Ethics and Recruitment

The Study received ethical approval from the South London and Maudsley SLaM, as well as from the Institute of Psychiatry Local Research Ethics Committee, Institute of Psychiatry, Psychology and Neuroscience (IOPPN)/South London and Maudsley (SLaM) research ethics approval number: 135/05. Informed written consent was obtained from all participants in the study by the GAP team.

As part of the GAP study (Di Forti et al., 2015, 2009). Patients aged 18–65 years who presented with first-episode psychosis at the inpatient units of SLaM were approached for recruitment. Patients were invited to participate if they met the ICD-10 criteria for a diagnosis of non-affective (F20–F29) or affective (F30–F33) psychosis (Guest et al., 2014), validated by administration of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Aboraya et al., 1989), they were re-contacted after the start of treatment. Between May 1, 2005, and May 31, 2011, 461 patients with first-episode psychosis were recruited. The cohort consisted of a diverse multi-ethnic population. Further patient information, blood samples and genetic ancestry were acquired as described previously (Di Forti et al., 2012). During the same period, a total of 389 control individuals were recruited. These individuals were aged 18–65 years and were similar to the local population regarding gender, ethnic origin, education, and employment status, and socio-economic status. Recruitment of controls was done using Internet and newspaper advertisements and by distributing leaflets at train stations, shops, and job centres. Volunteers were administered the Psychosis Screening Questionnaire (Bebbington and Nayani, 1996) and were excluded if they met the criteria for a psychotic disorder or if they reported a previous diagnosis of psychotic illness.

Gene Expression Data

Whole blood samples were collected using PAXgene tubes for RNA. Psychotic patients were stabilized using antipsychotics for a week before taking blood samples. Samples were run at the NIHR Biomedical Research Centre for Mental Health (BRC-MH) microarray facility at the MRC-SGDP, IOPPN, and King's College London. Microarrays were run following the manufacturer's protocol using Illumina HT-12 V4 beadchips (Illumina, USA), which contain approximately 48,000 probes.

The raw probe level gene expression data had been processed in GenomeStudio and exported for further analysis in R using the Lumi package (Du et al., 2008). The data was then analysed using the Illumina Gene Expression pre-processing pipeline developed by the BRC-MH Bioinformatics core (see Section 2.3 for more details).

Figure 2.1 provides an overview of how many samples were available at each stage of processing. In short out of 679 samples that were available to the microarray facility, 608 samples were provided after quality control in GenomeStudio. Only 480 (287 cases, 193 controls) of these samples fulfilled my initial inclusion criteria. This is because the initial 608 samples included 114 follow-up samples, as well as 14 samples representing mothers of patients. The preprocessing pipeline used network analysis for outlier detection, which resulted in the removal of an additional 97 samples (see Section 2.3.3 for details). It should be noted that while the RNA integrity number (RIN) was included in the analysis, we did not set a cut-off for it, instead it was incorporated into the pre-processing stage (see section 2.3.4). As a result 24 samples had a RIN below 8, with the lowest sample that passed our quality control pipeline having a RIN of 5.

After preprocessing, an additional 103 samples were excluded for a variety of reasons. A total of 34 samples were found to be duplicates, 31 samples could not be used since they could not be identified or had withdrawn consent and 5 samples had incomplete information for key variables (age, ethnicity, clinical gender). Finally 33 samples were removed since they were did not have First Episode Psychosis when the blood was taken.

This resulted in 280 samples (131 FEP and 149 controls) that were available for analysis. Chapter 3 provides analysis of the demographics from a case control perspective. The same samples are also used in Chapters 4 and 5, however the 131 FEP patients are for some analyses separated into Schizophrenia and Other Psychosis (based on ICD10 and DSMIV diagnosis). Table 2.1 provides an overview of demographic variables and medication for all 3 groups.

Genetic Data

Samples of patients and controls were genotyped using the Illumina HumanCore Exome BeadChip at the BRC Genomics Laboratory (South London and Maudsley NHS Trust, King's College London). Sample DNA was derived from either blood (80%) or saliva (20%). If both samples were available for an individual Blood was used. Genotypes were processed using Genome-studio Analysis (version 2011.1, Illuminia Inc.)

Genetic data was processed, and Polygenic Risk Scores were calculated by Evangelos Vassos as described in (Vassos et al., 2017). Calculation of PRS was performed using the Psychiatric Genomics Consortium (PGC2) data as the discovery sample, leaving out the

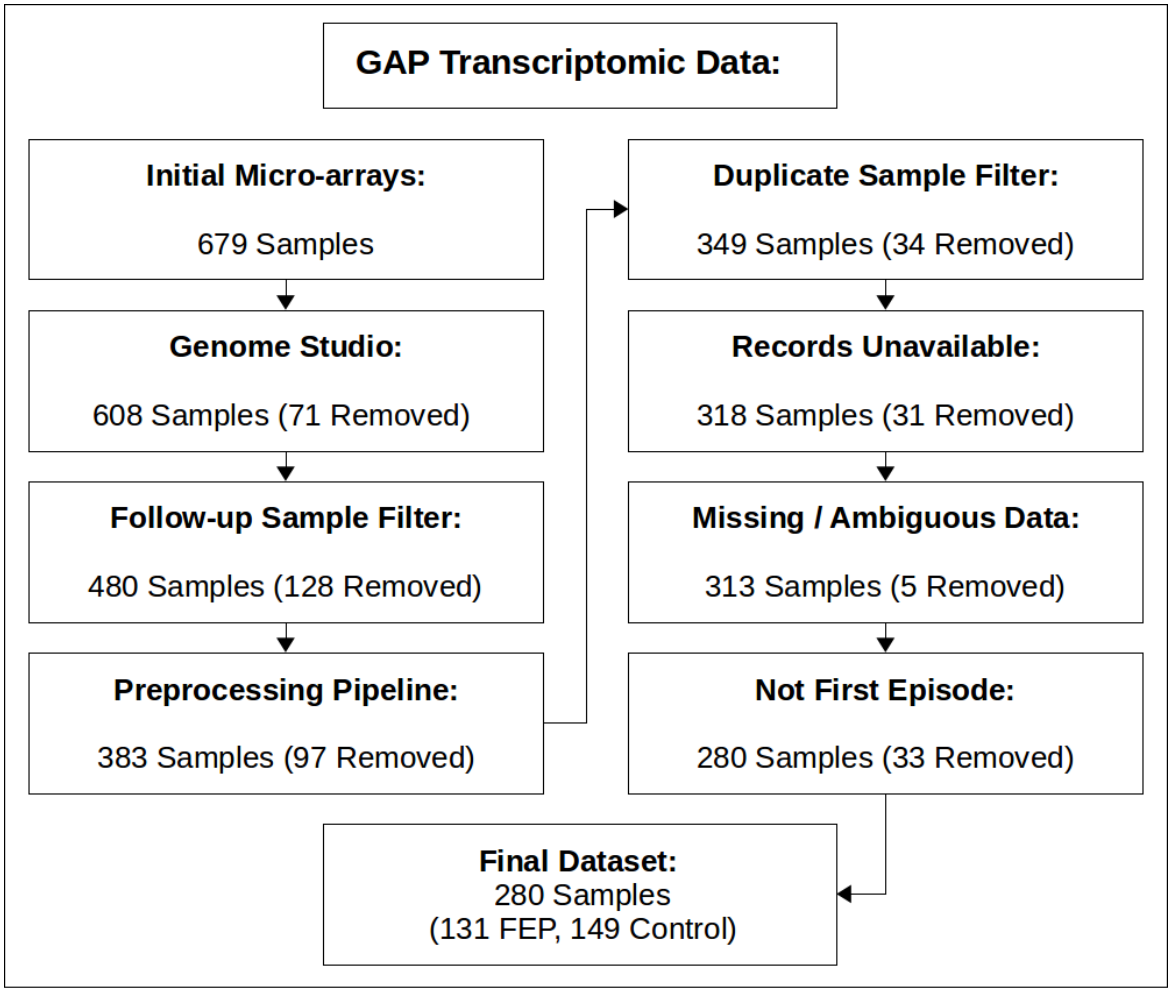


Figure 2.1 GAP Quality Control of Samples

Flowchart of samples that were removed during processing and experimental design. A total of 679 blood samples were used. After initial processing in genome studio 608 were suitable for further analysis. A total of 128 samples were excluded since they were blood samples from followup appointments. A further 97 samples were removed in the preprocessing pipeline (see Section 2.3), primary due to network based outlier detection. Finally a total of 103 samples were excluded for a variety of reasons related to their clinical records. This included 34 samples that were either duplicate entries, or had conflicting identifications. A further 31 samples could not be used due to clinical records not being available. Five samples were excluded due to ambiguous demographic data. Finally 33 samples had to be removed since they were identified as Chronic Psychosis, Non-organic Psychosis, or unconfirmed First Episode Psychosis.

Table 2.1 GAP Demographics

	Control	Other Psychosis	Schizophrenia
n	149	63	68
Gender = male (%)	86 (57.7)	31 (49.2)	45 (66.2)
Age (mean (sd))	29.87 (10.53)	30.03 (9.22)	26.59 (7.67)
Ethnicity (%)			
Asian	10 (6.7)	6 (9.5)	6 (8.8)
Black	43 (28.9)	22 (34.9)	33 (48.5)
Other	10 (6.7)	5 (7.9)	5 (7.4)
White	86 (57.7)	30 (47.6)	24 (35.3)
Medication (%)			
Amisulpride	0 (0.0)	1 (1.6)	0 (0.0)
Antipsychotic free	0 (0.0)	9 (14.3)	9 (13.2)
Aripiprazole	0 (0.0)	3 (4.8)	9 (13.2)
CONTROL	149 (100.0)	0 (0.0)	0 (0.0)
Haloperidol	0 (0.0)	4 (6.3)	2 (2.9)
Olanzapine	0 (0.0)	24 (38.1)	22 (32.4)
Quetiapine	0 (0.0)	2 (3.2)	4 (5.9)
Risperidone	0 (0.0)	13 (20.6)	14 (20.6)
Sulpiride	0 (0.0)	0 (0.0)	1 (1.5)
Trifluoperazine	0 (0.0)	0 (0.0)	1 (1.5)
NA	0 (0.0)	7 (11.1)	6 (8.8)

Table of GAP demographics by control, other psychosis and schizophrenia. Gender, Age, Ethnicity and Medication are listed for the 3 groups.

Welcome Trust Case Control Consortium 2 data since 80 of those samples were present in GAP (Psychosis Endophenotypes International Consortium et al., 2014). Genetic data were available for 445 FEP cases and 265 controls in GAP. We successfully identified 243 out of 280 samples as having both Gene expression and PRS data available. The first ten principal components were used to adjust PRS.

Missing Data

GAP presents a challenge due to its broad and complex nature. The cohort contains a significant amount of information far beyond a case-control study. Data were collected by multiple hospitals over more than a decade, and by dozens of physicians and clinical staff.

Combining clinical data with biological data was challenging, as not all patients gave consent for the same data, or were subject to different subsets of tests. While GAP has a wide variety of data on family history, psychological tests, assessments, clinical notes, and biological samples, this full range was rarely available for an individual. As such compromises had to be made in selecting data for analysis.

The work in this thesis included individuals with complete records for first episode psychosis, age, ethnicity and gender. In cases where additional data is required, such as Medication, BMI, PANSS, diagnosis (ICD-10 or DSMV) or Polygenic Risk Score, one of two strategies is used.

These are imputation approaches using machine learning techniques or literature-based predictions. Alternatively, subset analysis on the part of the cohort with full information is performed.

This is subject to a lot of researchers degrees of freedom and can introduce significant bias by testing many hypotheses and reporting only significant ones. To avoid this, the aim is to be transparent about all analysis that was performed posthoc.

No attempt was made to adjust for the number alternative hypotheses tested, but results that are the result of these analyses would need to be independently verified in separate datasets regardless.

2.1.2 Chronic Schizophrenia Data Set

The Chronic Schizophrenia data used for validation and machine learning is freely available at the ArrayExpress Archive (https://www.ebi.ac.uk/arrayexpress) with the identifier E-GEOD-38484. It was originally described by de Jong et al. (2012). In the original study, 239 samples were used from 2 platforms. Out of these 202 were from the Illumina H-12 microarray and 37 from Illumina H-8 microarray. The decision was made to exclude the 37 samples

from the Illumina H-8 array from analyses in this thesis, to avoid additional preprocessing and quality control steps, which could increase variation. We obtained the E-GEOD-38484 gene expression data using the ArrayExpress R package (Kauffmann et al., 2009). The gene expression data were mapped to corresponding reliably expressed probes in the GAP data where possible. Missing probes were substituted with probes corresponding to the same gene symbol.

2.2 Bio-informatics Methods

The main bioinformatic methods used in this thesis are differential gene expression (DGE), WGCNA and gene enrichment, and are described below. All analysis and programming was performed using R version 3.1.2 (Team, 2013), using R studio, unless otherwise specified.

2.2.1 Differential Gene Expression (DGE) Analysis

Standard differential gene expression (DGE) analysis was performed, using the R package LIMMA (Smyth, 2004) (version 3.26.8). Prior to analysis CellMix proportions, age, sex and ethnicity were regressed out. Probes were declared significantly differentially expressed if the false discovery rate (FDR) adjusted q-value was less than 0.05 and the absolute log fold change was above 0.1.

2.2.2 Weighted Gene Co-Expression Network Analysis (WGCNA)

The weighted gene co-expression network analysis (WGCNA) R package (Langfelder and Horvath, 2008) was used to generate gene expression modules. Duplicate probes mapping to the same gene were filtered out prior to analysis, since they tend to be highly correlated. This was done by comparing the average expression across all samples and keeping only the highest expressed probe for each gene for further analysis. An adjacency matrix was generated using a β of 6 which met the scale-free topology criteria. A hierarchical clustering tree was created and modules were originally defined using the WGCNA function cutreeDynamic with a minimum module size of 20. Modules were then merged using the mergeCloseModules function with a threshold value of 0.25. The eigengene of each module was then correlated with phenotypic information.

2.2.3 Gene Enrichment Analysis

All enrichment analysis was performed using the UserListEnrichment function in R (for details see https://cran.r-project.org/web/packages/WGCNA/index.html). This function is part of the WGCNA package (Langfelder and Horvath, 2008). Enrichment analysis, for the results of the LIMMA analysis, was performed by testing differentially expressed probes (q-value =< 0.05 and ± logFC > 0.1). All probes that did not pass the q-value threshold were included as background. Gene enrichment for probe lists corresponding to the modules identified in WGCNA analysis, was performed by using core genes of each module which were defined as probes with an above average module membership. All probes below this threshold, or belonging to other modules were labelled as background.

In all cases KEGG (Kanehisa et al., 2004) and GO (Ashburner et al., 2000) databases taken from the Enrichr website were used (http://amp.pharm.mssm.edu/Enrichr/stats, accessed 05.July.2016). The databases used were "KEGG 2016", "GO Molecular Function 2015", "GO Cellular Component" and "GO Biological Process 2015". Additionally, internal lists from the UserListEnrichment function in WGCNA were used. These were Brain Modules (useBrainLists), Brain Region Markers (useBrainRegionMarkers) and Blood Atlases (useBloodAtlases). Gene lists specifically compiled for Psychosis were included from publications by Purcell et al. (2014) and Pirooznia et al. (2016). Result categories that contained less than 5 overlapping probes were filtered out.

2.3 Transcriptomic quality control Pipeline

Transcriptomics has become a widely used technology in biology, which allows investigation of full expression in a tissue. This allows detection of transcriptional changes between time-points, tissues, diseases or environments. While this makes the technology extremely powerful and versatile, many factors must be taken into consideration, regarding experimental design, sample handling, data processing and downstream analysis.

Two key technologies are commonly used for transcriptomics, one of them is RNA-sequencing, which would have been cost prohibitive for this project, and the other is microarray technology based on hybridization. Gene expression microarrays fix, typically tens of thousands, complementary RNA to a microarray. Sample RNA is then labelled with fluorescent probes and can hybridise with the corresponding RNA on the array. This process allows high-throughput measurements of the relative levels of all included transcripts for a sample.

In this section, the processing of gene expression data for the thesis is explained.

2.3.1 Gene Expression Preprocessing Overview

The pre-processing of GAP gene expression data was performed using a modified version of the gene expression processing workflow developed by Dr Stephen Newhouse (Newhouse, 2013) for Illumina Beadarray data. The workflow contains the following five core steps.

1. Background Correction and Normalisation.
2. Outlier Detection.
3. Adjustment for confounding variables.
4. Test for expression of sex markers.
5. Detection of expressed probes.

2.3.2 Background Correction and Normalisation

Background correction in microarray gene expression experiments is a critical step in quality control, the Illumina BeadArray platform allows for background correction using negative non-specific control beads, of which thousands are present on each array. After careful consideration, we opted for three progressive steps which start with model-based background correction, followed by Log2 transformation and Robust Spline Normalisation (RSN).

This was based in part on analysis by Schmid et al. (2010) which compared a range of preprocessing methods and found that background correction in combination with Log2 and RSN processing provided the most accurate results across a multitude of measures. The background correction method used in that study was the standard suggested by Illumina which was previously reported not to make the best use of negative control beads(Barnes et al., 2005), and alternative methods have since been developed to address this, notably model-based background correction for BeadArrays (MBCB) (Ding et al., 2008) . The MBCB package (Xie, 2010) implements this and several other proposed methods within R as described by Allen et al. (2009). Of these, we use Maximum likelihood estimation.

Expression data was then further processed by using the lumiExpresso function from the Lumi package (Du et al., 2008). Background correction was disabled, and the log2 transformation was performed followed by Robust Spline Normalisation.

2.3.3 Network Analysis for Outlier detection

To detect outlying samples within both the control and psychosis group, a network-based approach was used. This was developed by Dr Stephen Newhouse for this pipeline and

is based on work by Oldham et al. (2012). This process involved generating networks of samples within each group based on expression data. Outliers were removed if standardised connectivity (Z.K) was below the threshold of -2. Z.K provides a measure of connectivity with the other samples in the group. A Z.K between -2 and -3 is recommended by Oldham et al. (2012) to identify members of the network that are substantial outliers. These samples are removed, and a new network is built on the remaining samples until all samples pass the Z.K threshold.

2.3.4 Correcting for technical and other confounding variables

Correcting for Batch Effects is essential in expression analysis, as any variation can affect reported expression values. It has been demonstrated that Batch effects when not properly accounted for can have a greater effect on the data than the underlying biological signal (Leek et al., 2010). To address this, we received available data on how samples were processed from the BRC-MH microarray facility at the Medical Research Council (MRC)-Social, Genetic & Developmental Psychiatry Centre (SGDP) (courtesy of Charles Curtis and Sang Hyuck Lee).

Since there is the possibility that batch effects can be introduced by unknown variables, Surrogate Variable Analysis SVA is used to identify hidden variables, via the Surrogate Variable Analysis (SVA) package (Leek et al., 2012). Of the known lab variables, we first removed highly correlated ones. Following this, we performed a principal component analysis on the gene expression matrix and extracted the first principle component. We identified four batch variables that were significantly correlated with the first principle component. The significant batch variables were "Date samples were thawed", "Concentration of initial RNA", "Concentration of Labelled cRNA", and "Date of RNA purification". It should be noted that while the RNA integrity number (RIN), was included in this step, it was not significantly correlated with the principle component. We therefore did not adjust for RIN, nor did we use a minimum cut-off for RIN.

We used multiple linear regression with the four batch variables as independent variables in the model. Residuals for all probes were extracted, and the average gene expression value of each sample was added to all probe residuals of that sample. Following this surrogate variable analysis was performed on the gene expression matrix, to capture hidden technical and biological variables that confound the data. As long as hidden variables are not substantially correlated with the underlying signal of interest (in this case Phenotype or Control vs First Episode Psychosis), SVA can be used to estimate hidden variables. This has been illustrated in simulated gene expression data (Leek and Storey, 2007), as well as in published data (Leek et al., 2010). SVA was performed on all datasets used in this thesis. Phenotype representing the binary labels for Control and Psychosis samples was

included in the model to preserve the biological signal between these groups. While in early iterations of preprocessing, hidden variables were identified, this was not the case in the final iteration of the pipeline used, or for publicly available data that had previously undergone preprocessing. Since SVA can potentially detect con-founders from any source, be it biological, environmental, or due to sample handling, this increases the confidence in our data and subsequent results.

The CellMix package (Gaujoux and Seoighe, 2013) makes use of some publicly available lists of marker genes for cell types to identify the likely composition of cells in a sample. These cell proportion estimates can be used in downstream analysis as covariates or regressed out as is described for technical variables above. The standard whole blood database by Abbas et al. (2009) was used to assess cell proportions with CellMix.

2.3.5 Expression based Sex detection

To detect potentially mislabelled samples, we performed a sex detection step based on X and Y chromosome expression. This was based on the X Inactive Specific Transcript (XIST) and Protein Kinase, Y-Linked, Pseudogene (PRKY) probes. XIST is usually only expressed in females, while PRKY is located on the Y chromosome and expressed in men. Samples that did showed aberrant expression for these probes were flagged, and additional records were individually examined. Initially, two samples were flagged, and they were ultimately excluded from the analysis.

2.3.6 Probe Selection

For each sample, probes were defined as expressed if background corrected mean expression for each probe was above the mean expression of negative beads, in the background corrected data, before normalisation. An additional step was performed in which only probes were kept which showed expression in more than 80% of sample per phenotype of interest (Control and Psychosis). We performed variations of this process to test reliability and found high agreement between methods (data not shown). This included the addition of Gender in the second step, resulting in four groups, by splitting Controls and Psychosis samples into two groups each, before detecting probes expressed in 80% of samples in those four groups. The probes identified as reliably expressed using this approach were then used in later analyses.

2.4 Machine Learning Methods

Machine Learning is a field that merges computer science and statistics with the aim of solving a problem in a way that has not been defined explicitly. Broadly speaking this means defining a task for a program and letting the program "learn" solutions from available data. The distinction between traditional statistics and machine learning is not always obvious, and definitions of what constitutes machine learning vary. In this thesis, a definition that is restricted to generating models capable of predicting categorical or numerical outcomes is used. This definition includes models built to predict a clinical diagnosis, or a model built to predict polygenic risk score. Methods that fall under this definition can range in complexity from Artificial Neural Network to simple Linear Regressions.

While these approaches can accomplish a broad spectrum of tasks, it involves several steps that need to be carefully considered to reach an optimal outcome. Appropriate preparation of data, selection of algorithms, training models, testing models and interpreting them correctly is vital.

In the following sections I will provide an overview of the steps involved in developing machine learning classifiers as well as the algorithms that were used in this thesis.

2.4.1 Data Preparation

Machine learning requires careful data preparation to generate predictive models. While this seems obvious, it turns out to be a complex topic. Some algorithms can handle varying amounts of missing, mislabelled or inaccurate information. Careful characterisation of predictive and outcome variables is therefore essential for all but the most trivial tasks.

It is also critical to consider how data is processed, how many variables are used and to what extent they are correlated. A large number of predictors (in this case RNA transcripts) that exceed the number of available samples for prediction (patients and controls) can easily lead to a classifier that identifies a unique signature for each sample by overfitting. Such a classifier may be highly accurate in the data it is trained on but would have no value in any other dataset. In addition, a large number of predictors can cause additional problems, by increasing processing time in a non-linear fashion, depending on the approach used.

This thesis is primarily based on gene expression data, with thousands of transcripts and only hundreds of samples, as such we use several methods to reduce the number of predictors. We define a threshold to remove the most highly correlated predictors, as two perfectly correlated predictors provide no additional information. We further remove predictors with the least variance, as they provide little information. The thresholds used for this are mentioned in the corresponding chapters. While this leads to a reduction of information

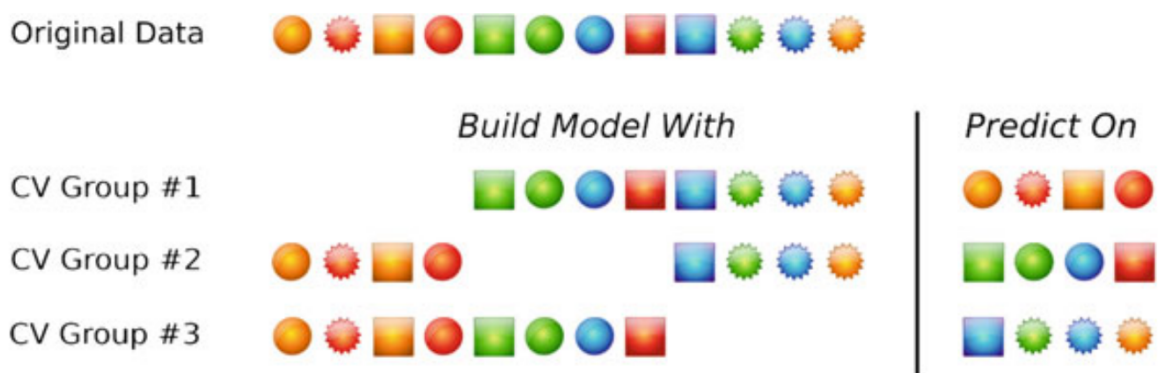


Figure 2.2 Cross-validation

Schematic representation of threefold cross-validation. The data is split into three thirds. Each third is iteratively left out from model building and the data is predicted on that group. *Republished with permission of Springer New York, from Applied Predictive Modelling by Johnson, Kjell; Kuhn, Max; permission conveyed through Copyright Clearance Center (Kuhn and Johnson, 2013).*

except for the most extreme examples, from a practical standpoint, this is a reasonable strategy.

2.4.2 Re-sampling Methods

To reduce the risk of over-fitting a variety of resampling methods can be used. The most common ones are cross-validation, repeated cross-validation and bootstrapping.

Cross-validation partitions the data into N number of equally sized sections. N number of models are then built, with each leaving out one of the predefined sections. The models are then evaluated on the left out data, and the results of all N models can be summarised to assess likely performance in new data. This process can be repeated for different tuning parameters, algorithms or predictors to find more robust combinations of all of them, before fitting a final model on the full data. See figure 2.2 for a schematic representation.

Repeated cross-validation is an extension of this that simply repeats this process several times, each time partitioning the data in a new way. Repeated cross-validation provides the best performance by balancing both bias and accuracy.

Bootstrapping functions differently to cross-validation. Where cross validation uses a subset of the overall to train, a bootstrap draws a random sample with replacement from the data (Efron and Tibshirani, 1986). This means that samples can appear multiple times in the training data. On average 63.2% of the samples are used in each bootstrap, meaning that 37.8% of the data used for training in each bootstrap iteration is duplicated. This also means that 37.8% of the data is on average not used for training, and can be used in testing later on. See figure 2.3 for a schematic representation.

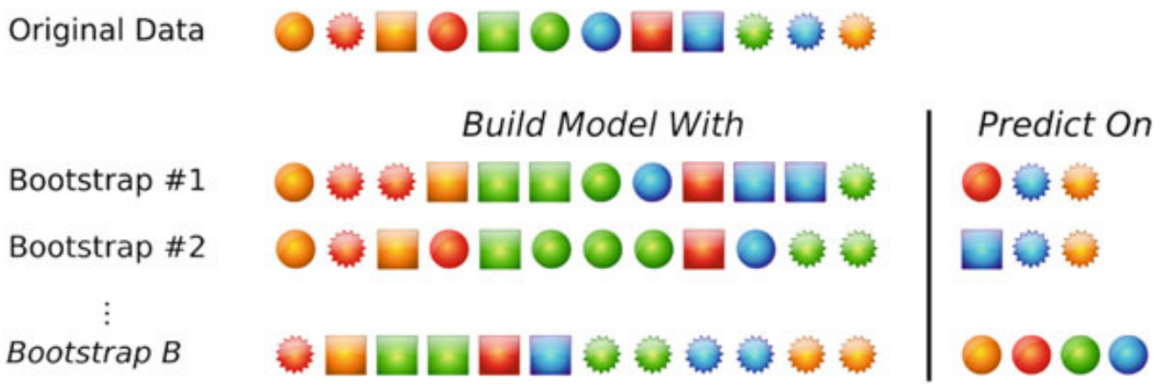


Figure 2.3 Bootstrap

Schematic representation of bootstrap resampling. Training data for the model is drawn with replacement from the original input data. Predictions are made on the samples that do not occur in the training data for the bootstrap iteration. *Republished with permission of Springer New York, from Applied Predictive Modelling by Johnson, Kjell; Kuhn, Max; permission conveyed through Copyright Clearance Center (Kuhn and Johnson, 2013).*

2.4.3 Generalised Linear Models with Regularisation (Glmnet)

Generalised linear models with regularisation are computationally efficient and well-defined (Hastie et al., 2009). This thesis makes use of the Glmnet package for R (Friedman et al., 2010) which offers efficient approaches for fitting generalised linear models with regularisation using lasso or elastic-net penalty. This is also covered by (Kuhn and Johnson, 2013) and integrated with the caret package. Glmnet has the advantage of being able to handle large numbers of correlated predictors and internally selecting predictors. This approach also provides estimated ranks of importance for the predictors in the final model. As such the GLMNET package provides an interpretable and efficient approach for initial machine learning experiments using gene expression data, where the number of predictors often outstrip the number of available samples.

2.4.4 K-Nearest neighbour (KNN)

K Nearest Neighbour is one of the simplest and most widely implemented(Altman, 1992) supervised learning methods. In simple terms, KNN predictions rely on calculating the distance between samples and assigning predictions based on the closest sample used to train the predictive model. The distance between samples can be defined in various ways depending on context, but the most common metric is Euclidean distance (Kuhn and Johnson, 2013). This means that processing of data, especially scaling as well as the number of predictors have a significant impact on this algorithm. The curse of high dimensionality

is a well-known problem in machine learning, and feature selection, as well as processing, becomes important (Aggarwal et al., 2001). We use KNN implementations via the caret package (Kuhn, 2008), for two purposes. One is for imputation of missing values based on a small number of predictors, and the other is as part of an effort to build ensemble classifiers from multiple classifiers built using different algorithms.

2.4.5 Naive Bayes (NB)

Naive Bayes represents a family of classifiers based on Bayes' theorem (Efron, 2013) , with a core assumption of uncorrelated predictors. Bayes' theorem is simply stated as follows:

$$Bayes'theorem : P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{2.1}$$

Where A is a categorical outcome and B is a predictor. The probability of observing the outcome or predictor is represented by $P(A)$ and $P(B)$ respectively. $P(B|A)$ is the probability of the predictor given an outcome and $P(A|B)$ is the probability of an outcome given the predictor. Naive Bayes algorithms use an assumption of independence between predictors to calculate the likelihood for all predictive values with the outcome variable. The likelihood of all predictors is ultimately merged to create predictions. While the predictors here are not independent, Naive Bayes can perform surprisingly well despite this assumption, and modern implementations report that additional processing can produce results that are comparable to those seen with support vector machines (Rennie et al., 2003). In this thesis we use the Naive Bayes implementation from the klaR (Weihs et al., 2005) R package, which is supported by the caret package (Kuhn, 2008).

2.4.6 Random Forests (RF)

Random Forests (Breiman, 2001) is one of the most widely used machine learning algorithms, due to the many packages that are available for implementation, high performance, protection from over-fitting, and ease of tuning and interpretation. A study of 179 popular machine learning packages across programming languages, found that Random Forests showed the best performance across 121 examined datasets. Three of the top five performing packages were RF implementations (Fernández-Delgado et al., 2014). Random Forests are an ensemble method that generates multiple decision trees and uses a random set of predictors for each tree. This process usually involves generating thousands of unique decisions trees with relative weak predictive power. The final prediction of a Random Forest for a sample is the

average, or majority prediction of all trees. The Random forest implementation in this thesis is via the randomForest package (Liaw and Wiener, 2002).

2.4.7 Support Vector Machines (SVM)

Support Vector Machines (SVM) was introduced by Cortes and Vapnik (1995) and have provided the best performance in predictive modelling until it became practical, due to computing power, to train finely tuned neural networks on large datasets. SVM algorithms are still widely used and performed as similar to random forests without additional fine-tuning by the user (Fernández-Delgado et al., 2014). SVM algorithms use kernel functions that are applied to available features. This approach allows for operations in higher dimensional space, which allows data that would otherwise not easily be separable with a hyperplane to be separated. For this thesis we use SVM with linear and polynomial kernels, implemented by the e1071 (Meyer, 2001) and kernlab (Karatzoglou et al., 2004) respectively. Both packages are integrated with the caret package (Kuhn, 2008).

2.4.8 Artificial Neural Networks (ANN)

Arguably the most powerful machine learning methods to date make use of Artificial Neural Networks (ANN) in some form or another. They were originally conceived in the 1940s and 50s by revelations about the animal brains interconnection of neurons and an interest in artificial intelligence. This led to the idea of perceptrons (Rosenblatt, 1958) which are mathematical representations of neurons. Simply put a single perceptron produces a binary response based on several input (I) values. Each input is assigned a weight (w), and if the sum of the weighted inputs exceeds a threshold (σ) the perceptron fires.

$$Perceptron: \text{output} = 1 \text{ if } \sum_{i=1}^{n} w_i I_i \geq \sigma \text{ else output} = 0 \tag{2.2}$$

Multiple perceptrons can be interconnected by creating layers that overlap in the input they receive or pass their output to other perceptrons, thus creating complex networks. This can quickly become very computationally intensive (Minsky and Papert, 1969), which prevented the use of Neural Networks until recently. In this thesis the pcaNNet from the caret package (Kuhn, 2008) was used, which is based on work by Ripley (1996). This implementation reduces computation time by using principal components to reduce the number of features.

2.4.9 Machine Learning Ensembles

Ensemble models are as the name suggests an ensemble of machine learning models, which aim to provide higher performance than the individual models. Ensemble approaches have in recent years led to the best performances in machine learning competitions, with bagging, boosting and stacking often being combined. However, they can be difficult to tune and implement in reality due to their complexity. Ensembles are often described in 3 subcategories. These are bagging, boosting and stacking.

Bagging or bootstrap aggregation is exemplified in random forests, where the ensemble is constructed from multiple models built on a random subset of the data. The resulting ensemble takes an unweighted majority vote of its constituent models to make predictions.

Boosting is an iterative process that builds an initial model, and emphasises prediction of misclassified data-points in following iterations. While this can lead to more accurate results, research indicates it leads to over-fitting and poor performance for supervised classification problems if there is a significant amount of uncertainty about the prediction labels. This is the case for mental health categories which are often highly heterogeneous and have high levels of uncertainty attached to them due to the subjectivity involved in making diagnoses. While boosting is used in this thesis, this is done within the implementing of the stochastic gradient boosting (gbm) package (Ridgeway, 2007) to train a stacking ensemble.

Stacking is an approach that combines the predictions of other algorithms. The Stacking algorithm can be trained with any number of algorithms and uses the predictions of previous algorithms as features. It is desirable to use uncorrelated predictive models for stacking ensembles since perfectly correlated model predictions do not add more information. Initial algorithms were therefore selected to represent different families of machine learning classifiers, which increases the likelihood of predictions being uncorrelated. The algorithms were also chosen for the high performance in the extensive analysis of machine learning algorithms on 121 datasets (Fernández-Delgado et al., 2014). The input models chosen here are glmnet, SVM (with linear and polynomial kernels), pcaNNet, nb and rf. Two SVM algorithms were included as the kernels significantly alter performance. In the case of neural networks, pcaNNet was used, as it uses principal components of available features to reduce dimensionality, and thus computational time. Stacking is implemented here using the caretEnsemble package (developer version) (Mayer, 2017). As mentioned above gbm is used to combine initial models built using the caret package for the final stacking ensemble in this thesis. This algorithm was chosen for the high performance and implementation of boosting. The rationale was the danger of over-fitting would be limited when using a handful of initial model outputs as predictors for each sample. Bootstrapping was used for the construction of input models, and the ensemble model.

Chapter 3

Differential Expression and Network Analysis

3.1 Introduction

In recent years genome wide association studies (GWAS) have resulted in a substantial advance in our understanding of the genetic components to Psychotic Disorders, such as Schizophrenia and Bipolar Disorders (Ripke et al., 2014). Much less focus, however, has been given to high-throughput gene expression analyses in the context of these disorders.

While complementary to GWAS, gene expression microarray studies have the advantage of not just analysing largely static genetic factors, but potentially reflecting dynamic responses to additional factors such as drug use, stress, age and other environmental factors. This is important since psychotic disorders are the result of a complex gene-environment interplay.

An important factor to consider when performing gene expression studies is the identification of a disorder relevant tissue. For pragmatic reasons, in this study, we chose to study transcriptional changes in whole blood, which is easily accessible and minimally invasive. There is an established literature of using blood for gene expression studies of a variety of psychiatric conditions, which include studies looking specifically at psychosis and schizophrenia. However sample sizes in this area have been small, ranging from dozens to about 100 patients (de Jong et al., 2012; Gardiner et al., 2013; Kumarasinghe et al., 2012; Kuzman et al., 2009; Lee et al., 2012; Wu et al., 2016). In addition few studies in this area are directly comparable, due to differences in microarray platform, and processing of results. In an attempt to address this Hess et al. (2016) recently pooled 9 brain, and 9 blood expression datasets for schizophrenia. However, it is as yet unclear if gene expression in first episode psychosis mirrors these expression patterns.

The aims of this study were, therefore, to identify genes with altered expression between first episode psychosis patients and controls. We performed a differential gene expression (DGE) analysis, followed by gene enrichment analysis and weighted gene co-expression network analysis (WGCNA). In addition, pathways associated with symptom severity were examined, by using PANSS. Finally, we examined the impact of medication.

3.1.1 Aims

Our aims for this study were as follows.

1. Explore differential expression between first episode psychosis cases and controls and tie this into the existing literature.
2. Find enriched pathways for differentially expressed probes.
3. Construct gene expression networks from the available data using WGCNA.
4. Find enriched pathways for WGCNA modules.
5. Correlate differentially expressed genes and modules with PANSS scores.
6. Estimate effect of antipsychotic medication on gene expression.

3.2 Methods

3.2.1 Gene Expression Data

The Genetics and Psychosis study was used in this analysis. Ethical approval, recruitment and detailed preprocessing are described in chapter 2.

In short, all analysis was performed using R version 3.1.2 (Team, 2013). We performed rigorous quality control, by pre-processing the data using an adapted in-house developed pipeline as described in chapter 2.

In short, the pipeline takes raw gene expression data exported from Illumina's Genomestudio, performs background correction (Xie et al., 2009) using negative bead expression levels to correct for noise. Lumi version 2.22.1 (Du et al., 2008) was used to log base 2 transform the data followed by robust spline normalization (Du et al., 2008). Outlying samples were iteratively identified using fundamental network concepts and removed, following the methods described by Oldham et al. (2012).

To reduce the influence of batch effects, we identified significant confounding variables by using the first principle component of housekeeping and undetected probes and regressing this against technical variables. In cases where the variables were significantly associated with the first principle component, they were regressed against expression for each probe, and the mean adjusted residuals were taken forward. The resulting adjusted expression matrix was subjected to Surrogate Variable Analysis, using the SVA package (Leek et al., 2012), to identify potential unknown batch effects. Following this, we compared recorded gender with gender determined by XIST and PRKY probes, and excluded samples that showed a mismatch. Finally, we excluded all probes that could not be reliably detected in 80% of the samples in at least one diagnostic group. We used the R package CellMix version 1.6 (Gaujoux and Seoighe, 2013), to test for potential significant differences in whole blood cell populations between cases and controls. Before further analysis we controlled for CellMix derived cell proportions, Age, Gender and Ethnicity using a linear model to create an adjusted expression matrix.

3.2.2 Linear Models for Microarray Data (LIMMA)

To identify differentially expressed genes (DE), the R package LIMMA (Smyth, 2004) (version 3.26.8) was used. Cell proportions, age, sex and ethnicity were previously regressed out. Probes were declared significantly differentially expressed if the FDR (false discovery rate) adjusted q-value was less than 0.05 and the absolute log fold change was above 0.1. Probes annotated with the “LOC” or ”HS.” prefix were filtered out at this stage.

3.2.3 Weighted Gene Co-Expression Network Analysis (WGCNA)

To identify modules based on co-expression we used the WGCNA R package (Langfelder and Horvath, 2008). For this analysis, we filtered out duplicate probes mapping to the same gene. An adjacency matrix was generated using a β of 6 which met the scale-free topology criteria. A hierarchical clustering tree was created, and modules were originally defined using the WGCNA function cutreeDynamic with a minimum module size of 20. Modules were then merged using the mergeCloseModules function with a threshold value of 0.25. The eigengene of each module was then correlated with phenotype information.

3.2.4 Gene Enrichment Analysis

All enrichment analysis was performed using the UserListEnrichment function in R (see https://cran.r-project.org/web/packages/WGCNA/index.html). This function is part of the

WGCNA package (Langfelder and Horvath, 2008). Enrichment analysis, for the results of the LIMMA analysis, was performed by testing differentially expressed probes (q-value =< 0.05 and ± logFC > 0.1). All probes that did not pass the q-value threshold were included as background.

For WGCNA enrichment analysis, this was done by using core genes of each module which were defined as probes with an above average module membership. All other probes were labelled as background. In all cases we used KEGG (Kanehisa et al., 2004) and GO (Ashburner et al., 2000) databases. We also used internal lists from the UserListEnrichment function in WGCNA, Brain Modules, Brain Region Markers and Blood Atlases. A list of Psychosis risk genes was taken from Pirooznia et al. (2016), this was an updated list, adapted from Purcell et al. (2009b). Result categories that contained less than five overlapping probes were filtered out.

3.2.5 Module correlation with Psychosis symptoms

Modules identified in the WGCNA were further correlated with the overall PANSS and three PANSS sub-scale (Positive, Negative and Psychopathology) scores. The gene expression matrices of modules associated with PANSS were scaled, centred and module genes were correlated with PANSS. Patients were stratified based on scores into low (n = 45, PANSS score = 7-15), medium (n = 29, PANSS score = 15-20) and high (n = 29, PANSS score = 20 - 49) symptom severity groups and gene expression for significant modules was plotted using heat-maps. Controls (n = 149) were used for comparison. Individuals with no available PANSS score were excluded (n=28).

3.2.6 Estimating effect of Medication

Differential expression analysis, as previously described, was used to test the effect of medication. Data was split into four groups based on medication status at the time of blood collection; all known Medication (Med), Olanzapine (Ola), Risperidone (Ris) and antipsychotic-free (AF). Four DGE analyses were performed, by using previously identified and significantly differentially expressed probes, with the following layout: 1. Med vs Control, 2. Ola vs Control, 3. Ris vs Control, 4. AF vs Control.

Antipsychotic data was also incorporated into WGCNA analysis. The groups were AF, Ola, and Ris. These groups were compared to controls by correlating them with WGCNA modules.

3.3 Results

3.3.1 Demographics

Of the 395 original samples (227 cases and 168 controls), 280 samples passed quality control and had full information on Age, Gender and Ethnicity. This corresponded to a final population of 131 first episode psychosis cases and 149 controls. The basic demographics of the final sample population are shown in Table 3.1. Patients were less likely to be Caucasian (p-value = 0.049) and more likely to be smokers (p-value =< 0.001). There was no significant difference in age, gender or body mass index.

Antipsychotic medication status at the time the blood was taken, is described below. Eighteen Patients (13.7%) were unmedicated when blood samples were taken, and information on medication for thirteen patients (9.9%) was unavailable. The remaining patients were primarily medicated with Olanzapine (35.1%), or Risperidone (20.6%), Aripiprazole (9.2%), Haloperidol (4.6%) and Quetiapine (4.6%). Trifluoperazine, Sulpiride and Amisulpride were all taken by a single individual.

Diagnosis of patients is detailed in Table 3.2 The most common ICD-10 diagnosis received by patients was schizophrenia (50.4%), followed by mania with psychosis (12.2%), schizoaffective disorder (6.9%), delusional disorder (3.8%), bipolar disorder (1.5%). In addition, various depression subtypes were diagnosed in 7.6% of patients, and 17.5% had incomplete records, did not meet any criteria, or had unspecified psychosis.

3.3.2 Differential expression

Of the 4730 probes remaining after quality control, 667 were removed due to being considered poorly annotated (LOC or HS. prefix), leaving a total of 4062 probes. In this reduced probeset we found 877 significantly DE genes (q-values < 0.05 & absolute logFC > 0.1). Out of these, 460 were up-regulated, and 417 probes were down-regulated (Appendix A: Table 1). The top 50 up and downregulated probes (q-values < 0.05) in terms of fold change are shown in Table 3.3. The most significant up and downregulated probes were, SUMO3 (logFC =0.1, q-value = 7.36E-05) and HNRNPUL2 (logFC =-0.18, q-value = 0.000206) respectively. The significant probes with the highest and lowest logFC were DEFA1B (logFC =0.91, q-value = 0.00021) and VWCE (logFC =-0.37, q-value = 0.00394) respectively.

Table 3.1 GAP Demographics

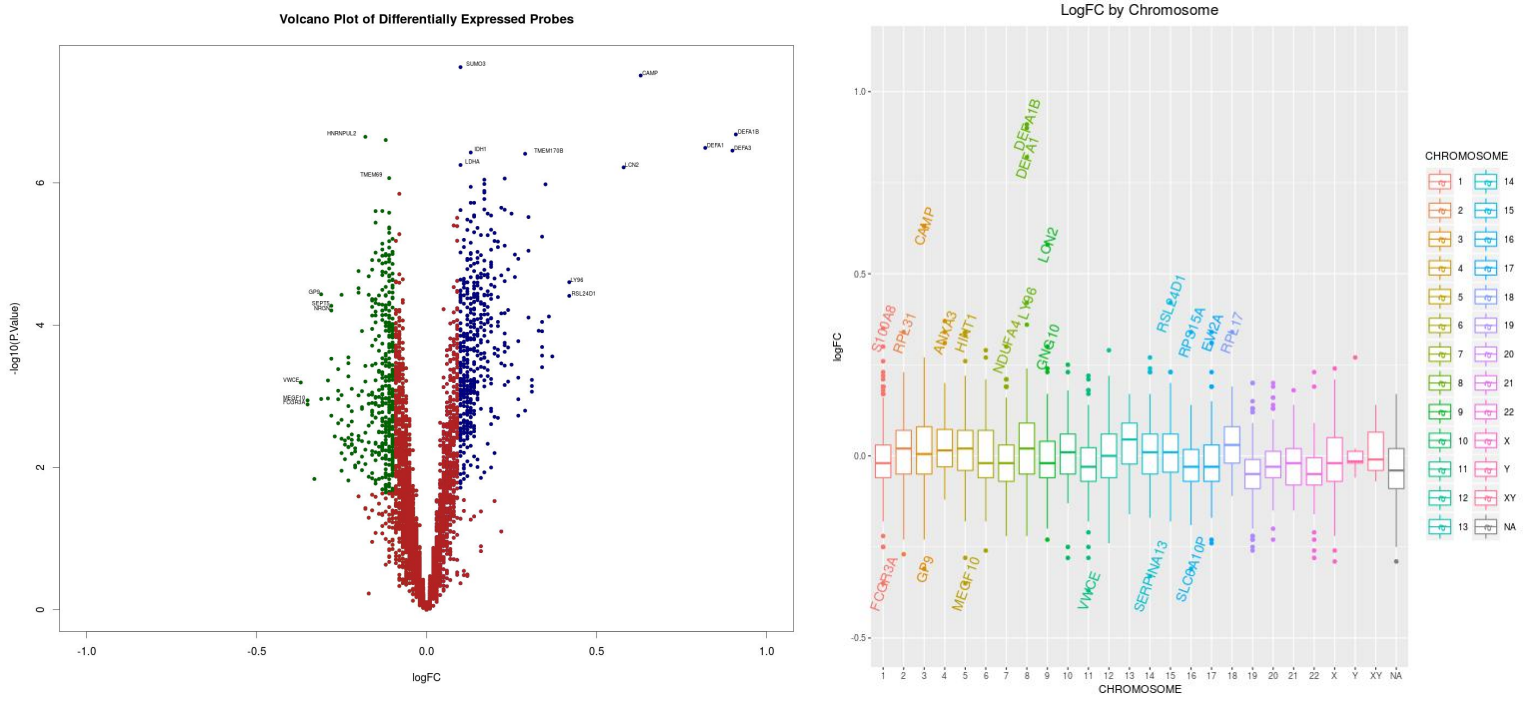
	Control	FEP	p-value
Samples	149	131	
Age (mean (sd))	29.87 (± 10.53)	28.24 (± 8.59)	0.163
Gender (%)			0.96
Male	86 (57.7%)	76 (58.0%)	
Female	63 (57.7%)	55 (58.0%)	
Ethnicity (%)			**0.049**
Asian	10 (6.7%)	12 (9.2%)	
Black	43 (28.9%)	55 (42.0%)	
Other	10 (6.7%)	10 (7.6%)	
White	86 (57.7%)	54 (41.2%)	
Tobacco use (%)			**<0.001**
Yes	79 (53.0%)	88 (67.2%)	
No	68 (45.6%)	30 (22.9%)	
NA	2 (1.3%)	13 (9.9%)	
BMI (mean (sd))	24.30 (± 4.55)	25.35 (± 5.59)	0.211

Table of Demographics for Controls and FEP patients. A total of 280 individuals were included in the study. The two groups showed significant differences in Ethnicity, with a higher proportion of White individuals in the control group. In addition Tobacco use was found to be more common in patients. No significant difference was found for Age, Gender or BMI. P-values for Gender, Ethnicity and Tobacco use were calculated using the chi-square test. Age and BMI p-values were calculated using the t-test.

Table 3.2 ICD-10 and DSM-IV diagnoses for Patients

ICD-10 (opcrit)	N (%)	DSM-IV (opcrit)	N (%)
Schizophrenia and delusional disorders			
Schizophrenia	66 (50.4)	Schizophrenia	32 (24.4)
		Schizophreniform Disorder*	27 (20.6)
Schizoaffective Disorder Manic	5 (3.8)	Schizoaffective Disorder Bipolar	7 (5.3)
Schizoaffective Disorder Depressive	4 (3.1)	Schizoaffective Disorder Depressed	9 (6.9)
Delusional Disorder	5 (3.8)	Delusional Disorder	7 (5.3)
Unspecified Non Organic Psychosis	5 (3.8)	Psychotic Disorder NOS	13 (9.9)
Affective Disorders			
Mild Depressive Episode	2 (1.5)		
Moderate Depressive Episode	2 (1.5)	Major Depressive Episode	2 (1.5)
Severe Depressive Episode with Psychotic Symptoms	6 (4.6)	Major Depressive episode with Psychotic features	11 (8.4)
Mania with Psychosis	16 (12.2)	Manic episode with psychosis	20 (15.3)
Bipolar Affective Disorder	2 (1.5)		
Unspecified			
Not Available	2 (1.5)	Not Available	2 (1.5)
No criteria Met	16 (12.2)	No Criteria Met	1 (0.8)

Table detailing the number of patients (N) followed by the percentage (%) of patients with the diagnosis. In total 131 patients are included. ICD-10 and DSM-IV diagnoses are based on the OPCRIT system (Rucker et al., 2011). *(All but 2 patients with Schizophreniform disorder are diagnosed with Schizophrenia under the ICD-10 criteria)



(a) Volcano Plot

(b) Breakdown by Chromosome

Figure 3.1 Visualizations of Differential Expression

(a) Shows a volcanoplot of limma differential expression results. Blue probes are up regulated in first episode psychosis, green probes are down-regulated. Red probes are considered unchanged either due to low p-value, or low differential expression. (b) Shows differential expression broken down by chromosome. The category XY refers to genes that are present on both the X and Y chromosome. The NA category contains probes that could not be mapped to single chromosome. Probes at the top are upregulated in psychosis patients, while probes towards the bottom are down-regulated.

Table 3.3 Top differentially expressed probes

Gene	logFC	p-value	q-value	Definition
Top 50 Up-regulated				
SUMO3	0.1	2.4e-08	7.4e-05	Homo sapiens SMT3 suppressor of mif two 3 homolog 3 (S. cerevisiae) (SUMO3), mRNA.
CAMP	0.63	3.1e-08	7.4e-05	Homo sapiens cathelicidin antimicrobial peptide (CAMP), mRNA.
DEFA1B	0.91	2.1e-07	2.1e-04	Homo sapiens defensin, alpha 1B (DEFA1B), mRNA.
DEFA1	0.82	3.2e-07	2.1e-04	Homo sapiens defensin, alpha 1 (DEFA1), mRNA.
DEFA3	0.9	3.5e-07	2.1e-04	Homo sapiens defensin, alpha 3, neutrophil-specific (DEFA3), mRNA.
IDH1	0.13	3.8e-07	2.1e-04	Homo sapiens isocitrate dehydrogenase 1 (NADP+), soluble (IDH1), mRNA.
TMEM170B	0.29	3.9e-07	2.1e-04	Homo sapiens transmembrane protein 170B (TMEM170B), mRNA.
LDHA	0.1	5.6e-07	2.4e-04	Homo sapiens lactate dehydrogenase A (LDHA), transcript variant 2, mRNA.
LCN2	0.58	6.1e-07	2.4e-04	Homo sapiens lipocalin 2 (LCN2), mRNA.
C9ORF72	0.23	8.7e-07	2.8e-04	Homo sapiens chromosome 9 open reading frame 72 (C9orf72), transcript variant 1, mRNA.
LYPLAL1	0.17	9.1e-07	2.8e-04	Homo sapiens lysophospholipase-like 1 (LYPLAL1), mRNA.
TFRC	0.17	1.0e-06	2.8e-04	Homo sapiens transferrin receptor (p90, CD71) (TFRC), mRNA.
S100A8	0.35	1.0e-06	2.8e-04	Homo sapiens S100 calcium binding protein A8 (S100A8), mRNA.
PCMT1	0.13	1.1e-06	2.8e-04	Homo sapiens protein-L-isoaspartate (D-aspartate) O-methyltransferase (PCMT1), mRNA.
BNIP2	0.17	1.3e-06	3.1e-04	Homo sapiens BCL2/adenovirus E1B 19kDa interacting protein 2 (BNIP2), mRNA.
SLC30A9	0.17	1.4e-06	3.1e-04	Homo sapiens solute carrier family 30 (zinc transporter), member 9 (SLC30A9), mRNA.
SLC44A1	0.17	1.7e-06	3.5e-04	Homo sapiens solute carrier family 44, member 1 (SLC44A1), mRNA.
TCEB1	0.13	1.9e-06	3.6e-04	Homo sapiens transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C) (TCEB1), mRNA.
VAMP7	0.14	1.9e-06	3.6e-04	Homo sapiens vesicle-associated membrane protein 7 (VAMP7), mRNA.
GLRX	0.22	2.2e-06	3.7e-04	Homo sapiens glutaredoxin (thioltransferase) (GLRX), mRNA.
CLNS1A	0.17	2.4e-06	3.7e-04	Homo sapiens chloride channel, nucleotide-sensitive, 1A (CLNS1A), mRNA.
FAM96A	0.23	2.4e-06	3.7e-04	Homo sapiens family with sequence similarity 96, member A (FAM96A), transcript variant 1, mRNA.
H2AFZ	0.1	2.4e-06	3.7e-04	Homo sapiens H2A histone family, member Z (H2AFZ), mRNA.
SENP7	0.25	2.7e-06	3.7e-04	Homo sapiens SUMO1/sentrin specific peptidase 7 (SENP7), transcript variant 2, mRNA.
COX7A2L	0.18	2.7e-06	3.7e-04	Homo sapiens cytochrome c oxidase subunit VIIa polypeptide 2 like (COX7A2L), nuclear gene encoding mitochondrial protein, mRNA.
C14ORF100	0.14	2.9e-06	3.7e-04	Homo sapiens chromosome 14 open reading frame 100 (C14orf100), mRNA.
TAF7	0.19	2.9e-06	3.7e-04	Homo sapiens TAF7 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 55kDa (TAF7), mRNA.
PSMC2	0.14	3.0e-06	3.7e-04	Homo sapiens proteasome (prosome, macropain) 26S subunit, ATPase, 2 (PSMC2), mRNA.
S100A12	0.3	3.0e-06	3.7e-04	Homo sapiens S100 calcium binding protein A12 (S100A12), mRNA.
MED28	0.12	3.3e-06	3.8e-04	Homo sapiens mediator complex subunit 28 (MED28), mRNA.
ARL6IP5	0.13	3.5e-06	3.9e-04	Homo sapiens ADP-ribosylation-like factor 6 interacting protein 5 (ARL6IP5), mRNA.
IFNGR1	0.15	4.3e-06	4.3e-04	Homo sapiens interferon gamma receptor 1 (IFNGR1), mRNA.
SRP9	0.19	4.6e-06	4.3e-04	Homo sapiens signal recognition particle 9kDa (SRP9), mRNA.
PRDX3	0.12	4.6e-06	4.3e-04	Homo sapiens peroxiredoxin 3 (PRDX3), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
ATG3	0.14	5.0e-06	4.4e-04	Homo sapiens ATG3 autophagy related 3 homolog (S. cerevisiae) (ATG3), mRNA.
CRLS1	0.2	5.1e-06	4.4e-04	Homo sapiens cardiolipin synthase 1 (CRLS1), mRNA.

Continued on next page

Table 3.3 – continued from previous page

Gene	logFC	p-value	q-value	Definition
CCPG1	0.16	5.4e-06	4.4e-04	Homo sapiens cell cycle progression 1 (CCPG1), transcript variant 2, mRNA.
CLDND1	0.18	5.5e-06	4.4e-04	Homo sapiens claudin domain containing 1 (CLDND1), transcript variant 1, mRNA.
ANXA3	0.34	5.7e-06	4.6e-04	Homo sapiens annexin A3 (ANXA3), mRNA.
TM2D1	0.11	6.3e-06	4.7e-04	Homo sapiens TM2 domain containing 1 (TM2D1), mRNA.
MAP2K1IP1	0.2	6.6e-06	4.7e-04	Homo sapiens mitogen-activated protein kinase kinase 1 interacting protein 1 (MAP2K1IP1), mRNA.
COX7A2	0.21	6.7e-06	4.7e-04	Homo sapiens cytochrome c oxidase subunit VIIa polypeptide 2 (liver) (COX7A2), mRNA.
FAM45A	0.13	7.0e-06	4.8e-04	Homo sapiens family with sequence similarity 45, member A (FAM45A), mRNA.
ATP5C1	0.23	7.6e-06	5.1e-04	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1 (ATP5C1), nuclear gene encoding mitochondrial protein, transcript variant 2, mRNA.
CDC40	0.11	8.0e-06	5.2e-04	Homo sapiens cell division cycle 40 homolog (S. cerevisiae) (CDC40), mRNA.
VBP1	0.21	8.0e-06	5.2e-04	Homo sapiens von Hippel-Lindau binding protein 1 (VBP1), mRNA.
PHF5A	0.19	8.7e-06	5.3e-04	Homo sapiens PHD finger protein 5A (PHF5A), mRNA.
GNG10	0.3	9.0e-06	5.3e-04	Homo sapiens guanine nucleotide binding protein (G protein), gamma 10 (GNG10), mRNA.
PSMD10	0.15	9.4e-06	5.3e-04	Homo sapiens proteasome (prosome, macropain) 26S subunit, non-ATPase, 10 (PSMD10), transcript variant 1, mRNA.
HEXB	0.1	9.5e-06	5.3e-04	Homo sapiens hexosaminidase B (beta polypeptide) (HEXB), mRNA.

Continued on next page

Table 3.3 – continued from previous page

Gene	logFC	p-value	q-value	Definition
Top 50 Down-regulated				
HNRNPUL2	-0.18	2.3e-07	2.1e-04	Homo sapiens heterogeneous nuclear ribonucleoprotein U-like 2 (HNRNPUL2), mRNA.
RBM14	-0.12	2.5e-07	2.1e-04	Homo sapiens RNA binding motif protein 14 (RBM14), mRNA.
TMEM69	-0.11	8.6e-07	2.8e-04	Homo sapiens transmembrane protein 69 (TMEM69), mRNA.
SCAP	-0.13	2.5e-06	3.7e-04	Homo sapiens SREBF chaperone (SCAP), mRNA.
FAM110A	-0.15	2.5e-06	3.7e-04	Homo sapiens family with sequence similarity 110, member A (FAM110A), transcript variant 3, mRNA.
RBM10	-0.11	2.6e-06	3.7e-04	Homo sapiens RNA binding motif protein 10 (RBM10), transcript variant 2, mRNA.
ZNF296	-0.15	3.6e-06	4.0e-04	Homo sapiens zinc finger protein 296 (ZNF296), mRNA.
PNPT1	-0.11	4.2e-06	4.3e-04	Homo sapiens polyribonucleotide nucleotidyltransferase 1 (PNPT1), mRNA.
RASGRP2	-0.11	4.4e-06	4.3e-04	Homo sapiens RAS guanyl releasing protein 2 (calcium and DAG-regulated) (RASGRP2), transcript variant 1, mRNA.
DENND4B	-0.11	5.0e-06	4.4e-04	Homo sapiens DENN/MADD domain containing 4B (DENND4B), mRNA.
ZC3H5	-0.1	6.1e-06	4.7e-04	PREDICTED: Homo sapiens zinc finger CCCH-type containing 5 (ZC3H5), mRNA.
CXXC1	-0.11	6.6e-06	4.7e-04	Homo sapiens CXXC finger 1 (PHD domain) (CXXC1), mRNA.
SUPT5H	-0.13	7.2e-06	4.9e-04	Homo sapiens suppressor of Ty 5 homolog (S. cerevisiae) (SUPT5H), mRNA.
GANAB	-0.11	8.1e-06	5.2e-04	Homo sapiens glucosidase, alpha; neutral AB (GANAB), transcript variant 2, mRNA.
PHF15	-0.13	8.7e-06	5.3e-04	Homo sapiens PHD finger protein 15 (PHF15), mRNA.
KIAA1267	-0.1	9.3e-06	5.3e-04	Homo sapiens KIAA1267 (KIAA1267), mRNA.
CLSTN1	-0.15	9.6e-06	5.3e-04	Homo sapiens calsyntenin 1 (CLSTN1), transcript variant 1, mRNA.
POM121C	-0.12	1.0e-05	5.3e-04	Homo sapiens POM121 membrane glycoprotein C (POM121C), mRNA.
TSSC4	-0.1	1.0e-05	5.3e-04	Homo sapiens tumor suppressing subtransferable candidate 4 (TSSC4), mRNA.
UBQLN4	-0.11	1.0e-05	5.3e-04	Homo sapiens ubiquilin 4 (UBQLN4), mRNA.
RANGAP1	-0.12	1.0e-05	5.3e-04	Homo sapiens Ran GTPase activating protein 1 (RANGAP1), mRNA.
OSBPL7	-0.13	1.1e-05	5.3e-04	Homo sapiens oxysterol binding protein-like 7 (OSBPL7), transcript variant 1, mRNA.
WDR23	-0.11	1.2e-05	5.6e-04	Homo sapiens WD repeat domain 23 (WDR23), transcript variant 1, mRNA.
FBXO46	-0.16	1.2e-05	5.6e-04	PREDICTED: Homo sapiens F-box protein 46, transcript variant 5 (FBXO46), mRNA.
PRKD2	-0.14	1.2e-05	5.6e-04	Homo sapiens protein kinase D2 (PRKD2), mRNA.
VAMP2	-0.11	1.3e-05	5.7e-04	Homo sapiens vesicle-associated membrane protein 2 (synaptobrevin 2) (VAMP2), mRNA.
NUMA1	-0.12	1.3e-05	5.8e-04	Homo sapiens nuclear mitotic apparatus protein 1 (NUMA1), mRNA.
ST3GAL1	-0.13	1.4e-05	5.9e-04	Homo sapiens ST3 beta-galactoside alpha-2,3-sialyltransferase 1 (ST3GAL1), transcript variant 1, mRNA.
EDC4	-0.1	1.4e-05	5.9e-04	Homo sapiens enhancer of mRNA decapping 4 (EDC4), mRNA.
SIK3	-0.11	1.6e-05	6.5e-04	Homo sapiens SIK family kinase 3 (SIK3), mRNA.
CD97	-0.16	1.7e-05	6.6e-04	Homo sapiens CD97 molecule (CD97), transcript variant 1, mRNA.
TLN1	-0.2	1.7e-05	6.8e-04	Homo sapiens talin 1 (TLN1), mRNA.
STIP1	-0.13	1.8e-05	6.8e-04	Homo sapiens stress-induced-phosphoprotein 1 (Hsp70/Hsp90-organizing protein) (STIP1), mRNA.
ITPKB	-0.12	1.8e-05	6.9e-04	Homo sapiens inositol 1,4,5-trisphosphate 3-kinase B (ITPKB), mRNA.
RAB35	-0.1	1.9e-05	7.0e-04	Homo sapiens RAB35, member RAS oncogene family (RAB35), mRNA.
RAB11FIP1	-0.13	2.0e-05	7.0e-04	Homo sapiens RAB11 family interacting protein 1 (class I) (RAB11FIP1), transcript variant 3, mRNA.
RASAL3	-0.11	2.0e-05	7.0e-04	Homo sapiens RAS protein activator like 3 (RASAL3), mRNA.
WASF2	-0.17	2.1e-05	7.0e-04	Homo sapiens WAS protein family, member 2 (WASF2), mRNA.
PHRF1	-0.11	2.1e-05	7.0e-04	Homo sapiens PHD and ring finger domains 1 (PHRF1), mRNA.

Continued on next page

Table 3.3 – continued from previous page

Gene	logFC	p-value	q-value	Definition
ICAM2	-0.1	2.3e-05	7.3e-04	Homo sapiens intercellular adhesion molecule 2 (ICAM2), transcript variant 1, mRNA.
HGS	-0.1	2.3e-05	7.3e-04	Homo sapiens hepatocyte growth factor-regulated tyrosine kinase substrate (HGS), mRNA.
MEF2D	-0.11	2.4e-05	7.3e-04	Homo sapiens myocyte enhancer factor 2D (MEF2D), mRNA.
SPG7	-0.1	2.8e-05	7.6e-04	Homo sapiens spastic paraplegia 7 (pure and complicated autosomal recessive) (SPG7), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
XRCC6	-0.12	2.9e-05	7.7e-04	Homo sapiens X-ray repair complementing defective repair in Chinese hamster cells 6 (XRCC6), mRNA.
FYN	-0.1	3.0e-05	7.8e-04	Homo sapiens FYN oncogene related to SRC, FGR, YES (FYN), transcript variant 2, mRNA.
PDPR	-0.2	3.0e-05	7.8e-04	PREDICTED: Homo sapiens pyruvate dehydrogenase phosphatase regulatory subunit (PDPR), mRNA.
TRIM28	-0.13	3.2e-05	8.1e-04	Homo sapiens tripartite motif-containing 28 (TRIM28), mRNA.
AKAP13	-0.11	3.4e-05	8.4e-04	Homo sapiens A kinase (PRKA) anchor protein 13 (AKAP13), transcript variant 2, mRNA.
AP1G2	-0.1	3.5e-05	8.4e-04	Homo sapiens adaptor-related protein complex 1, gamma 2 subunit (AP1G2), mRNA.
MED25	-0.2	3.5e-05	8.5e-04	Homo sapiens mediator complex subunit 25 (MED25), mRNA.

3.3.3 Enrichment of Differentially Expressed Genes

Enrichment analysis on all differentially expressed (DE) (N=877) genes revealed 97 categories at a p-value threshold of 0.05 (see Appendix A: Table 2). A total of 29 categories remained significant using a bonferroni adjusted p-value threshold of 0.05 (see Table 3.4). Enriched pathways (after accounting for multiple testing) included the Ribosome, Translation, RNA processing, Viral infections, protein transport and membrane targeting. The pathways contained between 23 and 83 probes and these overlapped significantly across categories, with most probes being members of the cytosolic or mitochondrial ribosome. Two of the significantly enriched categories (identified as Nucleus and Ribosome) are brain expressed modules that have been supplied by the WGCNA UserListEnrichment function.

Out of the remaining 68 categories that passed the p-value threshold of 0.05, but failed to reach a q-value threshold of 0.05, most fall into the above mentioned categories (see Appendix A: Table 2). The ones that did not included Innate immunity, glutamatergic synaptic function, the ubiquitin system, blood platelets, Parkinson's disease, exocytosis, learning or memory, nervous system development, the Amygdala, mir-137 and response to ethanol.

Table 3.4 Enriched Pathways for Differential Expression

Enriched Category	Library	Overlap	p-value	adj.p.val
Ribosome_Homo sapiens_hsa03010	KEGG_2016	43	0.0001	0.0001
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	GO_BP	43	0.0001	0.0001
viral transcription (GO:0019083)	GO_BP	36	0.0001	0.0001
cotranslational protein targeting to membrane (GO:0006613)	GO_BP	43	0.0001	0.0001
turquoise_M14_Nucleus_HumanMeta	Brain	83	0.0001	0.0001
protein targeting to membrane (GO:0006612)	GO_BP	48	0.0001	0.0001
establishment of protein localization to endoplasmic reticulum (GO:0072599)	GO_BP	44	0.0001	0.0001
protein targeting to ER (GO:0045047)	GO_BP	43	0.0001	0.0001
ribosomal subunit (GO:0044391)	GO_CC	42	0.0001	0.0001
translational termination (GO:0006415)	GO_BP	36	0.0001	0.0003
protein localization to endoplasmic reticulum (GO:0070972)	GO_BP	43	0.0001	0.0004
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)	GO_BP	41	0.0001	0.0006
structural constituent of ribosome (GO:0003735)	GO_MF	42	0.0001	0.0013
translational elongation (GO:0006414)	GO_BP	37	0.0001	0.0016
macromolecular complex disassembly (GO:0032984)	GO_BP	41	0.0001	0.0026
cellular protein complex disassembly (GO:0043624)	GO_BP	36	0.0001	0.0035
cytosolic large ribosomal subunit (GO:0022625)	GO_CC	23	0.0001	0.0042
large ribosomal subunit (GO:0015934)	GO_CC	25	0.0001	0.0052
protein complex disassembly (GO:0043241)	GO_BP	39	0.0001	0.0081
mRNA catabolic process (GO:0006402)	GO_BP	48	0.0001	0.0084
protein localization to membrane (GO:0072657)	GO_BP	54	0.0001	0.0093
establishment of protein localization to membrane (GO:0090150)	GO_BP	53	0.0001	0.0105
viral life cycle (GO:0019058)	GO_BP	38	0.0001	0.0121
nuclear-transcribed mRNA catabolic process (GO:0000956)	GO_BP	47	0.0001	0.0123
ribosome (GO:0005840)	GO_CC	40	0.0001	0.0139
translational initiation (GO:0006413)	GO_BP	41	0.0001	0.0301
RNA catabolic process (GO:0006401)	GO_BP	50	0.0001	0.0308
protein targeting (GO:0006605)	GO_BP	53	0.0001	0.0426
salmon_M12_Ribosome_HumanMeta	Brain	32	0.0001	0.0499

Table of results for gene enrichment analysis of different expressed genes between Controls and First Episode Psychosis. A total of 877 probes were differentially expressed. Results show categories that had a bonferroni adjusted p value below 0.05. (See Appendix A: Table 2 for full results)

3.3.4 Weighted Gene Co-Expression Network Analysis Results

WGCNA resulted in 14 modules after merging initial modules based on eigengene distance (Figure 3.2c). Network dendrograms of probes and module assignment before and after merging is shown in figure 3.2d.

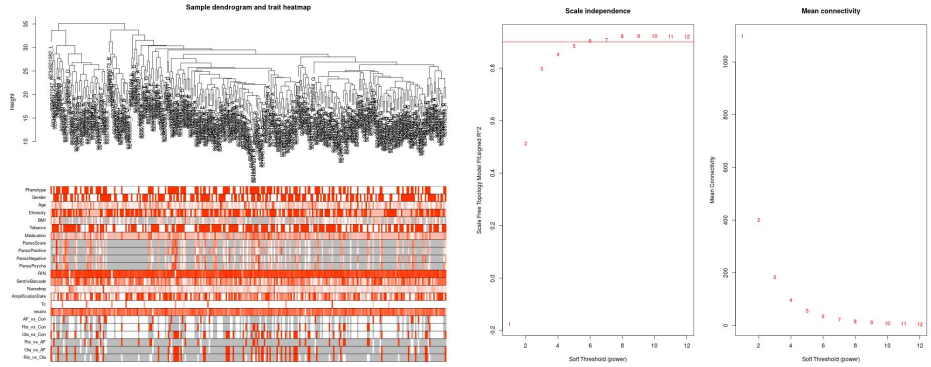
Module eigengenes were tested for their correlation with traits including case control status, BMI and Tobacco use. Correlations between controls status and three patient subgroups based on medication was also included. These were antipsychotic-free (AF) patients and patients on either Risperidone or Olanzapine. Figure 3.3 depicts a heatmap of these relationships, for each trait and module. Out of the 14 modules 7 were found to be statistically significant for the binary comparison between case and control. These were, in order of lowest to highest p-value the Blue (R^2 =0.26, p-value= $1x10^{-5}$, Size = 781 probes), Turquoise (R^2 =-0.24, p-value= $7x10^{-5}$, Size = 1663 probes), Cyan (R^2 =-0.22, p-value= $2x10^{-4}$, Size = 46 probes), Yellow (R^2 =-0.2, p-value= $6x10^{-4}$, Size = 338 probes), Green (R^2 =-0.2, p-value= $9x10^{-4}$, Size = 364 probes), Pink (R^2 =-0.15, p-value= $1x10^{-2}$, Size = 349 probes) and Tan (R^2 =-14, p-value= $2x10^{-2}$, Size = 69 probes) modules.

BMI was highly associated with the Pink, Purple, Magenta and Turquoise modules. Tobacco use was most significantly associated with the Cyan module. When comparing controls with three subsets of the FEP samples, namely just antipsychotic-free samples (AF vs HC), the antipsychotics Risperidone (Ris vs Con) or Olanzapine (Ola vs Con), we observed differences in several modules. Notably the Blue, Tan, Green and Turquoise modules. While the direction of the correlations stayed the same in all cases, the correlation for Risperidone was reduced in the Blue and Turquoise modules compared to antipsychotic-free and Olanzapine samples. No significant correlations were found between the Green module and Controls with Risperidone (p-value = 0.1) or Olanzapine (p-value = 0.03), but a significant negative correlation with the antipsychotic-free group was found compared to controls (R^2 = -0.23 p-value = $1x10^{-4}$). The Tan module showed the opposite pattern with the correlation being increased for Risperidone and Olanzapine compared to the antipsychotic-free group.

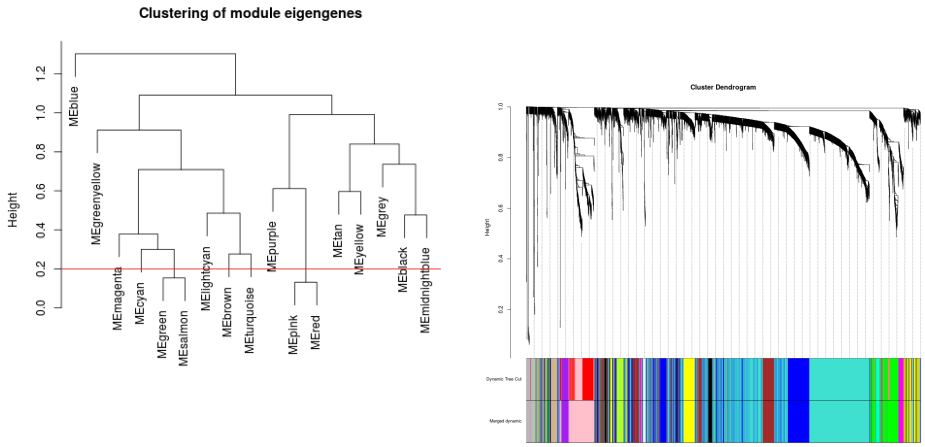
3.3.5 Enrichment analysis of WGCNA modules

For pathway analysis of the identified WGCNA modules, genes with above average module membership were used. The Grey module was excluded, due to being an aggregate of unassigned probes.

Enrichment analysis identified pathways in 8 of the 14 modules, using a q-value threshold of 0.05 and a minimum of 5 probes (see Table 3.5. The Pink, Purple, Midnightblue, Cyan,



(a) Clustering of samples and variables (b) Scale independence and connectivity



(c) Clustering of Eigengenes (d) Clustering of Probes

Figure 3.2 WGCNA Module Construction

(a) Hierarchical clustering of samples with (b) Scale free independence (c) Module eigengene hierarchical clustering with red line indicating height below which modules are merged. (d) hierarchical clustering of modules before and after merge with module membership indicated by colors.

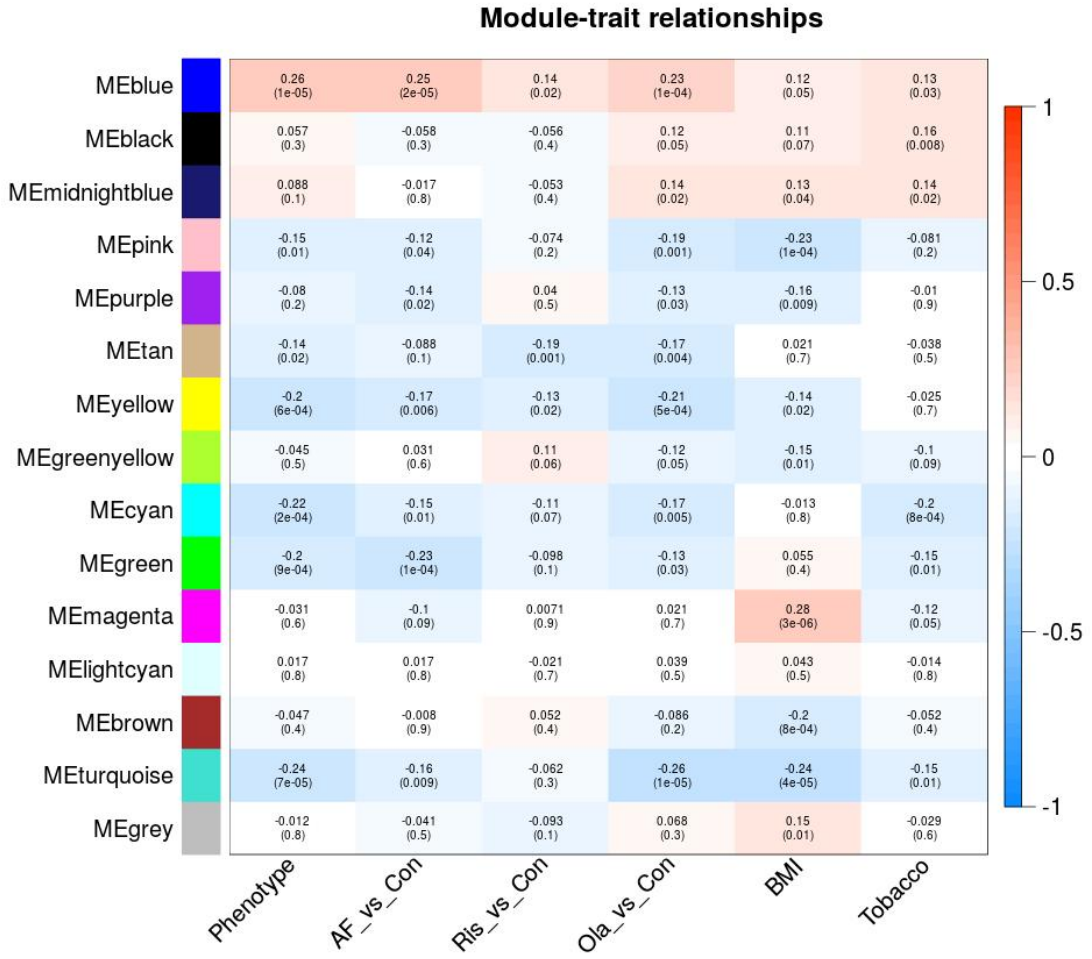


Figure 3.3 Module-trait Relationships

Heatmap of WGCNA modules and medication subgroups. The 5 categorical columns are coded as binary in correlations, with Control being 0. They are FEP vs. Control (Phenotype), Antipsychotic Free, Risperidone, Olanzapine. BMI is the body mass index, and Tobacco represents smoking status. Red signifies a positive correlation with the trait, while Blue represents a negative correlation. The top value is the correlation coefficient, while the bottom value is the p-value.

Green and Lightcyan modules did not contain modules that passed the multiple testing burden. Of the remaining modules, the Turquoise, Yellow, Blue and Green all had a correlation for case-control status above 0.2 and p-value below 0.001.

The most significant categories for the Turquoise Module was the Schizophrenia composite gene list adapted from Purcell et al. (2009b) “Scz-composite” (p-value = $4.52x10^{-18}$, q-value = $1.09x10^{-13}$), followed by ”UpWithAlzheimers_Blalock” (p-value = $1.15x10^{-11}$, q-value = $2.76x^{-7}$). For the Yellow module they were ”nucleolus” (p-value = 3.95x10-10, q-value = $9.52x10^{-6}$) and “Autism_Associated_Module_Vonieagu” (p-value = $1.1x10^{-6}$, q-value = 0.026). For the Blue module all 10 categories that passed the q-value threshold were brain derived modules. These included ”Glutamatergic Synaptic Function” (p-value = $2.06x10^{-7}$, q-value = 0.0049) and “DownWithAlzheimers_Blalock” (p-value = $3.01x10^{-7}$, q-value = 0.0072).

The Green module did not contain any enriched categories that passed the q-value threshold. The most significant category was “GlutatmatergicSynapse_Mouse” (p-value = 0.000186, q-value = 1).

3.3.6 WGCNA module relationship with Symptom Severity

PANSS scores were correlated with significantly differentially expressed probe and WGCNA modules (see Figure 3.4). The Greenyellow module was the most strongly associated module with Positive Symptoms (R^2 =0.29, p-value= $9x10^{-7}$). the correlation with negative symptoms was less significant and reversed (R^2 = -0.17, p-value= $5x10^{-3}$).

Positive Symptoms were also positively correlated with the brown (R^2 =0.15, p-value= $1x10^{-2}$), and blue modules (R^2 =0.17, p-value= $4x10^{-3}$), and negatively correlated with the black (R^2 = -0.19, p-value= $1x10^{-3}$), tan (R^2 = -0.2, p-value= $7x10^{-4}$), yellow (R^2 = -0.19, p-value= $2x10^{-3}$), green (R^2 = -0.18, p-value= $1x10^{-3}$) and grey (R^2 = -0.19, p-value= $1x10^{-3}$) modules.

The strongest association with negative symptoms was with the pink module (R^2 = -0.22, p-value= $2x10^{-4}$).

For the general psychopathology subscale associations were less significant, although there was a negative association with the green and tan modules. The strongest association for the overall PANSS was with the tan module (R^2 = -0.21, p-value= $6x10^{-4}$).

The probes in the Greenyellow module had expression levels scaled and centred and were stratified into four groups based on Positive (Figure 3.5) and Negative Symptoms scales (Figure 3.6). The low positive symptom group (PANSS = 15-20) had a general downregulation of almost all probe and high positive symptoms (PANSS = 20-49) correlated with increased expression. The reverse was true for Negative Symptoms.

Table 3.5 Enriched Pathways for WGCNA modules

Module	Library	Enriched Categories
Blue*	Brain	turquoise M14 Nucleus HumanMeta; turquoise M14 Nucleus MouseMeta;
Brain		yellow M18 CTX; blue M16 Neuron CTX; brown pyramidalNeurons Layer5/basolateralAmygdala Sugino/Winden; turquoise Cerebellum HumanChimp
Black	Brain	salmon M12 Ribosome HumanMeta
Transcription	GO BP	viral transcription (GO:0019083); translational termination (GO:0006415); cellular protein complex disassembly (GO:0043624); translational elongation (GO:0006414); translational initiation (GO:0006413); viral life cycle (GO:0019058)
	KEGG 2016	Ribosome Homo sapiens hsa03010
	GO CC	ribosomal subunit (GO:0044391); ribosome (GO:0005840); cytosolic part (GO:0044445)
	GO MF	structural constituent of ribosome (GO:0003735)
Tan*	Blood	Lymphcytes genesCorrelatedAcrossIndividuals Whitney
Yellow*	GO CC	nucleolus (GO:0005730); nucleoplasm (GO:0005654)
Brain	Brain	Autism associated module M12 Voineagu
Greenyellow	Brain	orange M5 Microglia(Type2) CTX
Brain	HBA	principal sensory nucleus of trigeminal nerve localMarker(top200) IN Pontine Tegmentum; Supraoptic Nucleus localMarker(FC>2) IN Hypothalamus
	GO BP	cellular response to type I interferon (GO:0071357); type I interferon signaling pathway (GO:0060337); response to type I interferon (GO:0034340); defense response to virus (GO:0051607); response to virus (GO:0009615); cytokine-mediated signaling pathway (GO:0019221)
	KEGG 2016	Herpes simplex infection Homo sapiens hsa05168; Measles Homo sapiens hsa05162; Hepatitis C Homo sapiens hsa05160; Influenza A Homo sapiens hsa05164
Magenta	Blood	Reticulocytes genesCorrelatedAcrossIndividuals Whitney; RedBloodCell Kabanova
Brown	Brain	pink M10 Microglia(Type1) HumanMeta
Brain	HBA	Substantia Nigra, pars reticulata localMarker(FC>2) IN Mesencephalon; Substantia Nigra, pars reticulata localMarker(top200) IN Mesencephalon
	Blood	RedBloodCell Kabanova
	GO CC	plasma membrane region (GO:0098590)
	GO BP	phagosome maturation (GO:0090382); extracellular matrix organization (GO:0030198); extracellular structure organization (GO:0043062)
	KEGG 2016	Osteoclast differentiation Homo sapiens hsa04380; Tuberculosis Homo sapiens hsa05152
Turquoise*	Pirooznia	Scz-composite; synaptome; neuronal PSD
Brain	Brain	UpWithAlzheimers Blalock ADvsCT inCA1; PostSynapticDensity proteins Bayes; Up CD40 stimulation in MG AitGhezala MicrogialMarkers; pink M14 GlutamatergicSynapticFunction CTX; green M5 Mitochondria HumanMeta; blue downAD metalIonTransport glycoprotein Blalock AD
	GO BP	actin filament-based process (GO:0030029); hemostasis (GO:0007599); blood coagulation (GO:0007596); coagulation (GO:0050817); actin cytoskeleton organization (GO:0030036); platelet activation (GO:0030168)
	GO MF	actin binding (GO:0003779); RNA polymerase II regulatory region DNA binding (GO:0001012)
	GO CC	extracellular vesicular exosome (GO:0070062); cell-cell junction (GO:0005911); microtubule (GO:0005874)
	KEGG 2016	Viral carcinogenesis Homo sapiens hsa05203; Endocytosis Homo sapiens hsa04144; Focal adhesion Homo sapiens hsa04510

Table of WGCNA modules and corresponding gene enrichment results. Criteria for inclusion of Enrichment Category were a q-value below 0.05 and a minimum of 5 probes overlapping with the pathway. Only 8 modules contained any Pathways that passed the inclusion criteria. Only probes with above median module membership were used for enrichment analysis. *Modules with an association to Psychosis status (p-value < 0.05)

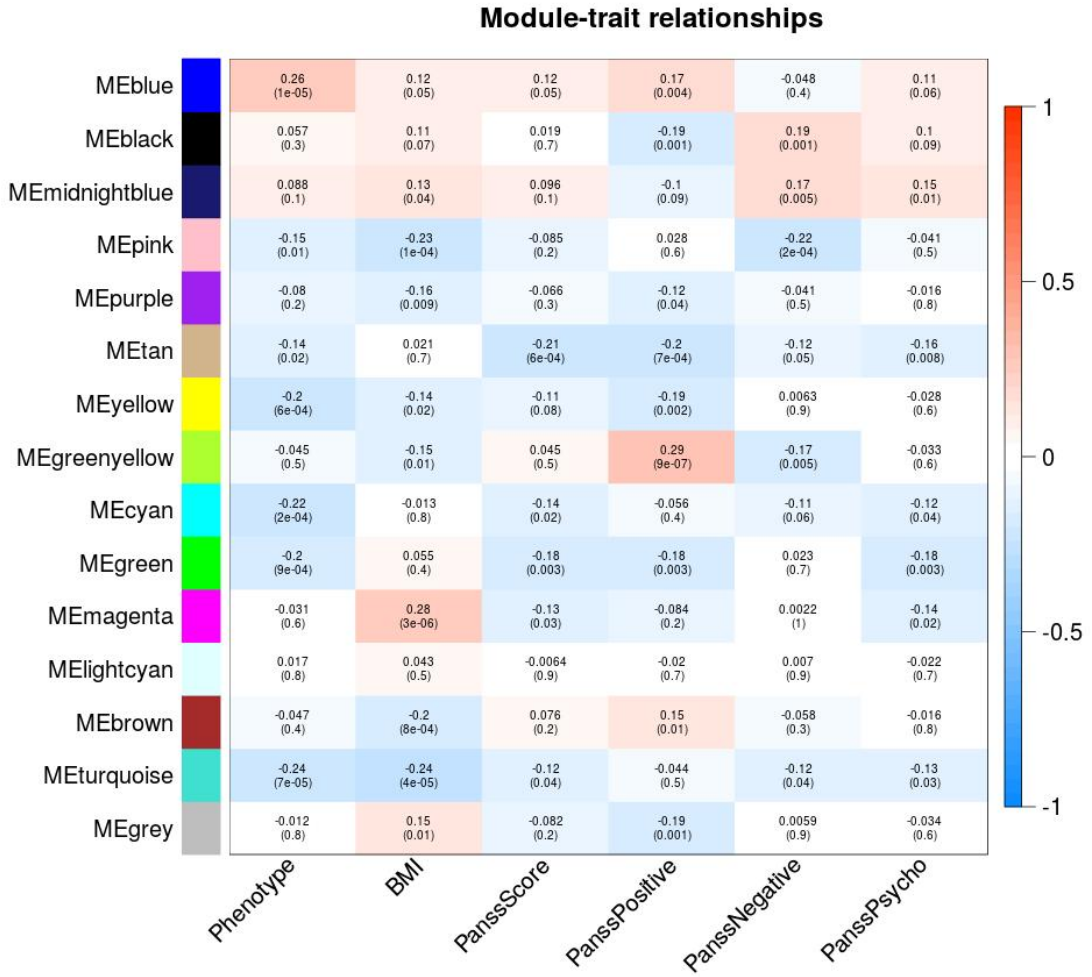


Figure 3.4 Module-trait Relationships PANSS

Heatmap of WGCNA modules and correlations with PANSS sub-scales. The first column represents FEP vs. Control coded as binary in correlations as before. The PannsScore is the overall PANSS, while PanssPositive, PanssNegative and PanssPsycho represent the 3 subscales. Red signifies a positive correlation with the trait, while Blue represents a negative correlation. The top value is the correlation coefficient, while the bottom value is the p-value.

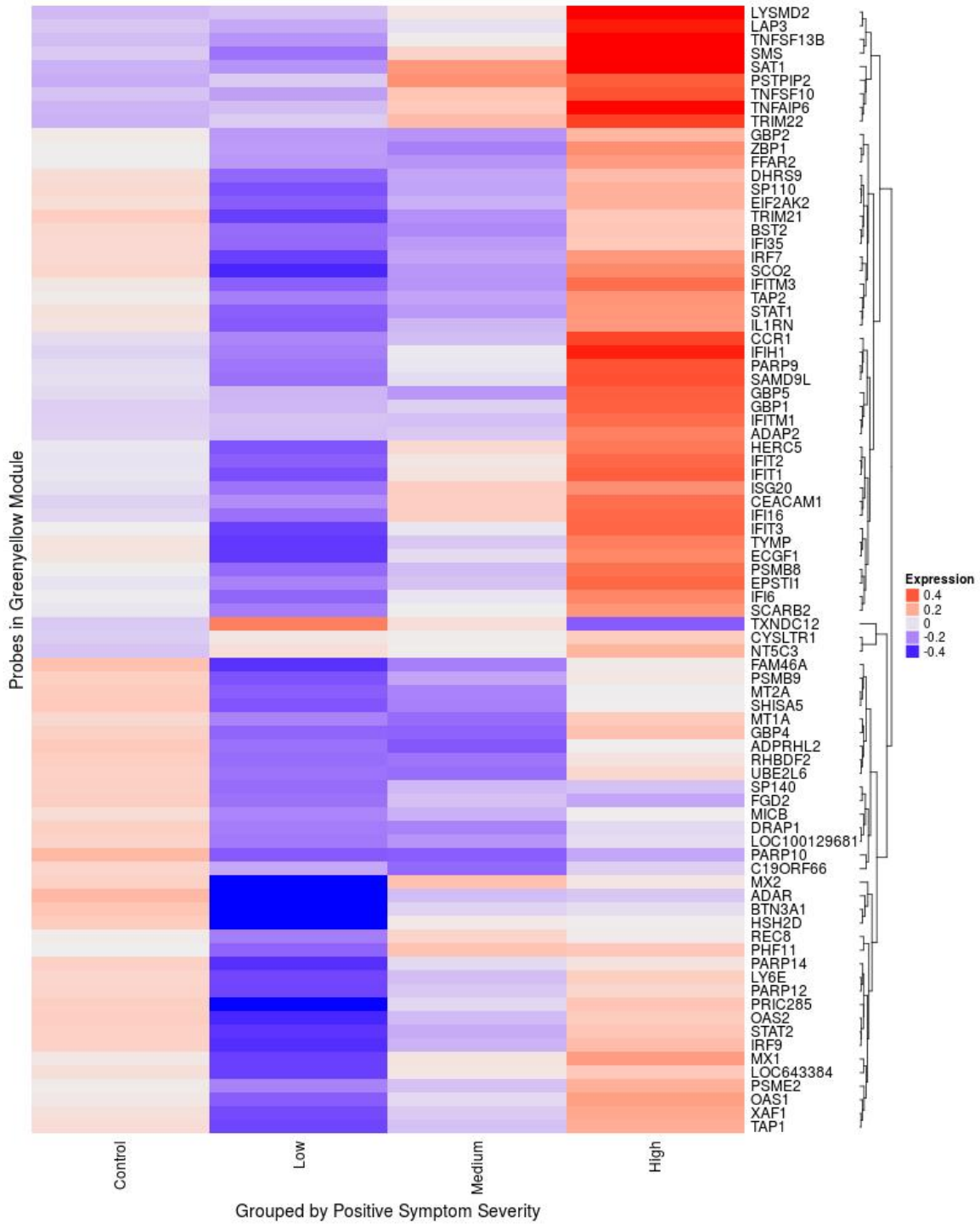


Figure 3.5 Heatmap of Expression Level in Greenyellow Module by Positive Symptoms

Heatmap of Probes in Greenyellow module based on positive symptoms. Rows correspond to indicated probes. Columns represent average expression level across individuals in each group. The groups were generated based on the score on the PANSS positive sub scale. Cases were split into low (n = 45, PANSS score = 7-15), medium (n = 29, PANSS score = 15-20) and high (n = 29, PANSS score = 20 - 49) positive symptom groups. Controls (n = 149) were used for comparison. Individuals with no available PANSS score were excluded. Rows are clustered using complete-linkage clustering, with the dendrogram indicating distance.

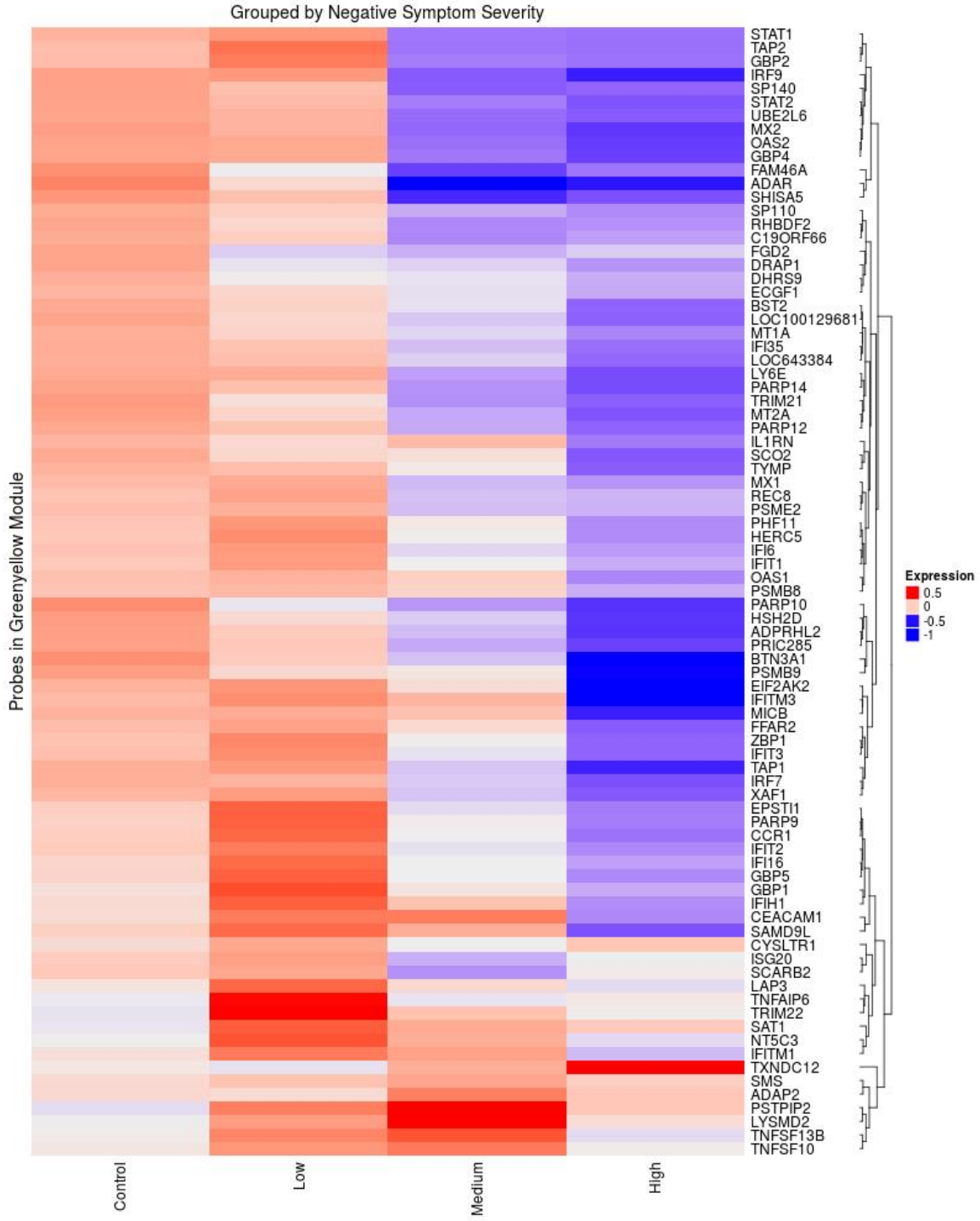


Figure 3.6 Heatmap of Expression Level in Greenyellow Module by Negative Symptoms

Heatmap of Probes in Greenyellow module compared with Negative symptoms. Rows correspond to indicated probes. Columns represent average expression level across individuals in each group. The groups were generated based on the score on the PANSS Negative sub scale. Cases were split into low (n = 45, PANSS score = 7-15), medium (n = 29, PANSS score = 15-20) and high (n = 29, PANSS score = 20 - 49) negative symptom groups. Controls (n = 149) were used for comparison. Individuals with no available PANSS score were excluded. Rows are clustered using complete-linkage clustering, with the dendrogram indicating distance.

3.3.7 Medication

To characterise the effect of antipsychotics in our sample, we used probes that were previously shown to be differentially expressed and performed secondary differential expression analysis on four subsets of the data based on medication. Healthy controls were compared with all samples with known medication (n=100), Olanzapine (n=46), Risperidone (n=27) and antipsychotic Free (n=18) groups (see Appendix A: Figures A.1 to A.4). No significant probes were found in the Risperidone group. All medication groups had a reduced log fold change across probes, while log fold change for probes were amplified when the antipsychotic free group was compared to controls.

3.4 Discussion

3.4.1 Differential Expression Pathways associated with Innate Immunity

When visualising differential expression, we notably see upregulation of Alpha Defensin probes in psychosis patients (Figure 3.1a), which include the three highest upregulated probes in FEP patients in this study (DEFA1, DEFA1B, DEFA3). This is consistent with a previous microarray study of similar size looking at schizophrenia patients (Gardiner et al., 2013) as well as two proteomic studies of Schizophrenia (Plasma) and Bipolar (Saliva) patients (Craddock et al., 2008; Iavarone et al., 2014). More recently they were also identified in a transcriptome meta-analysis by Hess et al. (2016), in a core module. Alpha Defensins are primarily involved in the innate immune system specifically in viral opsonisation. Upregulation of innate immune genes has been a consistent finding in the schizophrenia literature, and viral infections have been linked to increased schizophrenia risk, especially during pregnancy. In line with this, we also found the Cathelicidin Antimicrobial Peptide (CAMP) to be the second most upregulated probe in psychosis (logFC = 0.63, q-value = 7.36x10-05). These probes were not assigned to any module but were highly correlated among themselves.

The enrichment results overwhelmingly show changes in translation/transcription and mitochondrial function, in addition to the immune system, in line with previous results found by Hess et al. (2016) and others. Interestingly there is much overlap in the probes linked to viral replication and the above categories in our gene lists, which might point towards immune deregulation in psychosis patients. However such changes in the expression of innate immunity are also seen in chronic stress and sleep disorders which are well-known factors in schizophrenia and psychosis (Anderson and Maes, 2012). This is also consistent with cortisol deregulation via the HPA axis, which has been found in first episode psychosis

(Karanikas et al., 2014), and individuals at high risk of developing psychosis (Perkins et al., 2005).

While not passing the multiple testing thresholds, several of the categories that passed a p-value threshold of 0.05 were related to brain function. These included learning and memory, Parkinson's disease and glutamatergic synapses. Since only a subset of genes is expressed in both blood and brain tissues, brain-related pathways have a lower prior probability of being enriched in this analysis and are thus mentioned. They may provide insights, since Pathways that are deregulated in one blood may also be disrupted in the brain, either due to medication, drugs, hormones, sleep or genetic vulnerability.

3.4.2 Modules associated with Psychosis and Psychosis Severity

Using WGCNA several modules were identified that were significantly correlated with first-episode psychosis. These were the Blue, Yellow, Cyan, Green and Turquoise modules (see Figure 3.3). The Blue, Yellow and Turquoise modules were significantly enriched for multiple pathways, using a q-value threshold of 0.05, and are discussed further in this section.

While neither the Cyan or Green modules showed significant enrichment for any pathways using a q-value threshold, the Green module was enriched for pathways at a p-value threshold of 0.05, with the top category being related to glutamatergic neuron function. The Green module was therefore also included for discussion below.

Correlation with PANSS further highlighted the Greenyellow module, where an association with positive symptoms and was found (Figure 3.4). It was also highly enriched for brain-expressed genes and is discussed below.

Blue module

The Blue module was highly enriched for pathways from the Brain library of pathways (see UserListEnrichment documentation), with overlap for Amygdala and Cerebellum modules. This module was the only one that was positively correlated to psychosis. Interestingly when the Risperidone group was compared to controls, no significant correlation was observed, which stood in contrast to Olanzapine and antipsychotic-free patients.

Differential expression analysis, comparing these patient groups with controls, identified no significant probes for Risperidone, suggesting that Risperidone may have a major role in normalising gene expression patterns in blood, even after a short period of treatment. Olanzapine may be more specific to the brain or simply act on other pathways.

The Amygdala is involved in emotion regulation, and facial recognition and has been linked to schizophrenia as well as a series of other neuropsychiatric disorders. If this pathway

can be normalised using Risperidone identifying patients with deregulation in it, may be useful for targeted treatment.

However, we cannot exclude the possibility that this signal discrepancy may be due to differences in how Risperidone and Olanzapine are prescribed.

Turquoise module

The Turquoise module was enriched for brain pathways and schizophrenia risk genes. Using gene enrichment the category most strongly associated with this module was the previously curated list of genes by Purcell et al. (2014) of plausible schizophrenia genes. This represented the single most significantly enriched category across all modules.

The Hub gene with the highest absolute module membership for this module was Histidine Triad Nucleotide Binding Protein 1 (HINT1) (-0.95). HINT1 has previously been shown to interact with the Cannabinoid-1 Receptor (CNR1) to negatively regulate NMDAR activity (Vicente-Sánchez et al., 2013), and research by Di Forti et al. (2009) has provided evidence that cannabis increases the risk of developing schizophrenia sevenfold in at-risk populations.

The Turquoise module was also enriched for pathways associated with Actin processes, coagulation, vesicle transport and CNS associated pathways such as Glutamate, the Synaptome and Alzheimers Disease. Interestingly the module was most strongly correlated to the Olanzapine subset and BMI. Olanzapine is associated with significant weight gain, which can produce significant changes in gene expression. Correlation with Risperidone was again non-existent, indicating a potential Risperidone mediated effect in normalising gene expression levels to those of controls. This is especially interesting when taking into consideration the enrichment of schizophrenia genes.

However, since this module was by far the largest in the analysis with 1663 probes results should be regarded with caution.

Green Module

The Green module was also the only module that showed an increased correlation between antipsychotic-free samples and controls, compared to both medication subsets, or the overall case-control correlation.

Interestingly the two medication categories had no significant correlation with this module when compared to controls, indicating the module was similar in medicated patients and control individuals. This may mean that both Risperidone and Olanzapine act on this module.

This module also contained the only “GRIN” probe, namely Glutamate Ionotropic Receptor NMDA Type Subunit Associated Protein 1 (GRINA). These genes either are part

of Glutamate receptors or are associated with them. From our data in isolation, it seems like the Green module is upregulated in response to antipsychotics. The hub gene with the highest membership for this module was CLK2, a kinase that plays a role in gluconeogenesis and AKT1 dephosphorylation via PPP2R5B.

Further investigation with gene enrichment analysis revealed 6 of the top 10 modules to be associated with Protein signalling and degradation via the ubiquitin system, while the remaining four modules related to the nervous system. Of these two specifically mapped to Glutamate related pathways. While these passed the p-value threshold, non passed the q-value threshold. Also, the signal was based on less than 20 probes, and the green module is among the larger ones we identified, with 364 members. Nonetheless, I feel it is appropriate to highlight this module since pathway analysis is itself somewhat limited in that its based on characterised lists of genes. It is therefore conceivable that pathways associated with this module have not been defined in the literature, and given the prior probability of glutamate's role in schizophrenia, it is important to highlight this module.

Yellow modules

A previous study by de Jong et al. (2012) generated WGCNA modules and identified one (Tan) that was associated with chronic schizophrenia patients and replicated in non-medicated samples. It contains 129 probes, 108 of which were present in our dataset. The Yellow module (338 probes) presented here contains 40 of the 108 probes which are in the Tan module generated by de Jong et al. (2012). This supports the assertion of the authors that the overall pathway is robust and preserved. Interestingly the Yellow module is enriched for brain-expressed probes and genes previously associated with schizophrenia (Pirooznia et al., 2016), as well as an autism pathway. de Jong et al. (2012), found ABCF1 to be the HUB gene of the tan module they identified, and this is mirrored in our results were ABCF1 is a hub gene in the Yellow module. Also, a series of other genes specifically discussed by de Jong et al. (2012) were assigned to the yellow module including TUBB, RING1 and HSP90AB1 which the authors suggest are all linked to the MHC locus.

Greenyellow Module

Correlation of PANSS with WGCNA modules highlighted the greenyellow module. The highest correlation was with positive symptoms (0.29, p-value < 0.001), but the correlation with negative symptoms was also significant (-0.17, p-value = 0.005).

The Greenyellow module was enriched for pathways such as type I interferon signalling, viruses (including herpes) and three brain modules relating to Microglia, Principle sen-

sory nucleus of the trigeminal nerve (Pontine Tegmentum) and the Supraoptic Nucleus (Hypothalamus).

The herpes simplex virus, type 1 (HSV-1) is the primary cause of cold sores, and is known to hijack the trigeminal nerve and copy itself into DNA of the pontine tegmentum. The Hypothalamus is adjacent to this region, and they play an important role in REM sleep. HSV-1 has been associated with worse schizophrenia outcomes, by causing encephalitis in rare cases and attacking the temporal lobe which can lead to psychosis, problems with memory, behavioural changes and sleep problems (Prasad et al., 2012; Yolken, 2004). The Hypothalamus is also deeply involved in stress responses, via the HPA axis and cortisol. Activation of this axis can occur through psychological and physiological stress.

Surprisingly, individuals with low positive PANSS scores were related to the most significant signature, with expression being lower than for controls across almost all probes, while high scores on the positive symptom scale increased expression over controls. This was probably related to the negative correlation between positive and negative symptoms, as negative symptoms severity decreased global expression levels for the module.

The most strongly downregulated probe in high negative symptoms was Adenosine Deaminase, RNA Specific (ADAR), which was also, to a lesser extent downregulated in high positive symptoms when compared to controls. ADAR is a ubiquitously expressed gene, with an important role in RNA editing, the primary function of which is marking and differentiating host RNA from viral RNA. It is also linked disorders of the innate immune system, cancer and neurological disorders (Mannion et al., 2015).

ADAR mutations can cause severe autoimmune reactions and Aicardi-Goutieres syndrome, a neurodevelopmental disorder (Rice et al., 2012). In addition ADAR is located on chromosome 1q21.1, which was identified by the International Schizophrenia Consortium as a rare deletion increasing Schizophrenia risk (International Schizophrenia Consortium, 2008). It has been suggested that this may in part be due to its role in sleep regulation via the Glutamate system (Robinson et al., 2016), and due to its role in editing the serotonin receptor 5-HT2C mRNA (Yang et al., 2004). This is due to intracellular pattern recognition receptors for RNA stimulating the immune response via interferon signalling.

ADAR has been proposed to be required to suppress the innate immune systems interferon pathway from targeting internal RNA, by modifying host RNA to avoid an autoimmune responses. If this function breaks down, intracellular receptors like RIG-1 and IFIH1 (also known as MDA5), become activated and start their signalling cascades leading to transcription of Interferons, IL-6, iNOS and TNF-α via TRAF3, TBK1 and IRF-3/7 signalling (Wang et al., 2017). IRF-7 and IFIH1 are in both members of the greenyellow module, and they most strongly upregulated in strong positive symptoms. These results were also largely

mirrored by Hess et al. (2016) who found significant upregulation of TNF-α, Interferon, TLR, STAT, NF-κB and IL-6 related pathways. In addition to cellular stress responses such as Hypoxia, RNA metabolism, and Apoptosis. In reality, the distinction between the innate immune and cellular stress responses is arbitrary, as they are both involved in the stress response via HPA-axis activation. and similar findings were reported for gene expression studies in schizophrenia, depression and PTSD (Breen et al., 2017; Gardiner et al., 2013; Jansen et al., 2016).

Most of the genes in the greenyellow module were not differentially expressed between cases and controls. However, more than half of them were correlated above 0.25 with positive PANSS score. This indicates that this module may play a role in positive and negative symptoms as they are negatively correlated.

To test if a medication was the primary driver of this effect, an further analysis was performed looking at the correlation of positive symptoms in antipsychotic-free individuals (n= 13). The most significant probe in the greenyellow module became ADAR with a correlation of 0.72 with positive symptoms and -0.35 with negative symptoms.

The Greenyellow module provides evidence for HPA-axis deregulation, most likely due to increased stress, either by an increased exposure to life stressors (be they psychological, or mediated by pattern recognition receptors) or an increased sensitivity to stress. Either results in upregulation of the innate immune system and can cause an inflammatory response that is also transmitted to the brain. Pro-inflammatory cytokines such as Interferon type 1, and TNF can then affect behaviour by disturbing the serotonin and dopamine systems among others (Miller and Raison, 2016). This can also lead to microglia activation and oxidative stress in the brain, leading to behavioural changes, such as hyper-vigilance and anxiety. Accumulating damage to neurons and social rejection during disease progression, could in this way lead to paranoia and delusional thinking.

3.4.3 Effects of Medication

Differences were found in correlations between the three medication subgroups, especially when comparing Risperidone (N=27) to controls. This was evident from differential expression results which indicated a reduction in log fold change and identified no significant probes (see Appendix A: Figure A.3). It is possible that differences simply failed to reach statistical significance. However, almost all probes in the antipsychotic-free group were significantly differentially expressed, despite the smaller sample size (N = 18) (see Appendix A: Figure A.4). The Olanzapine (N = 46) comparison also identified many significantly differentially expressed probes, although the log fold change was much lower than for probes in antipsychotic-free samples (see Appendix A: Figure A.2).

The WGCNA results also suggested changes for the Risperidone vs control comparison in multiple modules in contrast with Olanzapine, antipsychotic-free and the overall FEP group comparisons with controls. A reduction in correlation for Risperidone was identified in the blue, pink, purple, yellow, cyan, green and turquoise modules.

This might be due to differences in demographics of the sub-populations, prescription tendencies or the time scale of effect of the drugs. Some studies have indicated that Olanzapine has a more tolerable side effect profile, and is more effective in the treatment of negative symptoms (Shoja Shafti and Gilanipoor, 2014), which may explain the more pronounced effect of Risperidone. The difference might also be due to individuals with negative symptoms receiving more Olanzapine. There was moderate evidence for this, with the average negative symptom score for the Olanzapine group being 17.5, while it is 15.3 for the Risperidone group.

Given the changes in log fold change in differential expression and the WGCNA data, it seems reasonable to conclude that the effect of antipsychotics is likely to change expression in the direction of the control group, meaning that any differentially expressed probes have a low likelihood of being significant purely because of medication.

Overall, these results are exploratory, and in need of further validation, unless they have prior support from other studies.

3.4.4 Conclusion

This chapter presented the results of the GAP gene expression data for the first time. Differential expression, WGCNA and gene enrichment analysis were used. Symptom severity and medication were explored in relation to WGCNA modules, and potential confounding variables such as Age, Gender, Ethnicity, Smoking and BMI were included when possible.

I found evidence for differential expression of probes involved in innate immunity, the viral response, metabolism and translation/transcription. Two of the significant pathways were expressed in brain. In addition I found subthreshold pathways relating to memory, the glutamate system and Parkinson's disease. While these pathways are worth pointing out they should be viewed with scepticism.

Using the WGCNA package, 15 modules were generated. Five of these have been discussed in detail. These are the blue, turquoise, yellow, green and greenyellow modules. All of these are enriched for brain-expressed genes. The results suggested that the blue and turquoise modules were affected by Risperidone, normalising expression towards controls. The turquoise module was highly enriched for schizophrenia risk genes. The yellow module partially replicated previous results in chronic schizophrenia.

The green module was the only module that correlated more strongly with the antipsychotic-free group than both Risperidone and Olanzapine groups, indicating it may be mechanically involved in antipsychotic function. The most significantly enriched pathway was related to glutamatergic neuronal functions, at a p-value threshold, but not a q-value threshold.

The greenyellow module was associated with positive and negative symptoms in opposite directions and was highly enriched for interferon gamma 1 signalling pathways. This supports previous literature showing links with the immune system in schizophrenia and other psychiatric disorders. This is discussed in the context of the Greenyellow module, and the hub gene ADAR.

We report that our differential expression analysis partially replicates and expands on previous findings of a series of psychiatric studies. The strongest signal is in association with the innate immune signal, transcriptional changes and mitochondrial perturbations. All of this is consistent with chronic stress via activation of the HPA axis, which leads to over-activation of the innate immune system, and ultimately changes to behaviour and brain function. This highlights the importance of taking stress responses more seriously. Especially since outcomes for mental health in developed countries have lagged behind those in the developing world, and genetic and molecular strategies have been at best been disappointing.

Despite this we show evidence of plausible gene expression signatures for psychosis in blood, and while these results need to be replicated they are highly encouraging for follow up studies.

Chapter 4

Predictive Modelling using the GAP data

4.1 Introduction

Accurate diagnosis is crucial to facilitate better patient care. Currently, in the case of psychosis, diagnosis is performed by clinical assessment and is somewhat subjective. There is a broad range of causal factors for psychosis ranging from drugs, sleep deprivation, physical diseases to schizophrenia and other psychiatric conditions. This requires identification of the underlying cause of psychosis for treatment to be most effective. Diagnostic concordance between the ICD-10 and DSM (III, IIIR, IV) categorisations lies between 71% and 93% of the DSM diagnoses showing higher concordance with the ICD-10, primarily due to Schizophreniform disorder and other subtypes, that fall under Schizophrenia within the ICD-10 (Jakobsen et al., 2006). In addition to this, diagnostic assessment of the same individuals using the same classification system also shows high discordance. This pattern is similar for other psychiatric conditions that feature psychotic symptoms, and newer versions of the DSM have radically increased diagnostic categories, largely for practical clinical purposes, rather than for distinct underlying biological, social or psychological reasons.

While there are certain cases of Psychosis in which the optimal treatment can be identified with relative ease, (for example drug abuse, NMDAR encephalitis) the majority of cases do not fall into a clear diagnostic category, and often go through several diagnoses, or receive additional ones during their lifetime. As such, finding biomarkers and other predictive clinical variables could potentially provide value by ruling out or identifying psychosis subtypes more readily. This could help guide treatment plans by identifying contributing causal factors of psychosis. In addition finding early markers of psychosis could help prevent relapse, predict response to medication or reduce the chance of developing psychosis to begin with, by allowing early intervention. The GAP study has multiple data-streams available, including genetic, transcriptomic, demographic and clinical data. As such it provides an opportunity to

evaluate the individual and combined predictive power of these data sources, in first episode psychosis and in ICD or DSM based subtypes.

4.1.1 Aims

In order evaluate the predictive power of these data-streams, we had 4 core aims:

1. The first was to build Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET) using various combinations of Gene expression data, polygenic risk scores and demographic information.
2. The second aim was to compare performance across models in first episode psychosis.
3. The third aim was to investigate feature importance for various models.
4. The fourth aim was to investigate how often samples were accurately classified, using bootstrapping and how diagnosis and symptom severity (PANSS) correlated with this.

4.2 Methods

4.2.1 Gene Expression Data

Available data consisted of 280 samples, with 149 controls and 131 first episode cases. We used four types of data in these experiments. The gene expression data was adjusted for cell type using CellMix (Gaujoux and Seoighe, 2013), as described in chapter 2 This full set of features was used as the Gx set. We also used the 1311 probe subset, consisting of genes with a hypothesized connection with schizophrenia. These genes were sourced from Pirooznia et al. (2016). We further made use of demographic variables. In this case, they consisted of Age, Sex and Ethnicity, and they are referred to as demographics or demo.

4.2.2 Polygenic Risk Score (PRS)

The polygenic risk scores (PRS) and principal components were calculated by Evangelos Vassos as described in (Vassos et al., 2017). PRS was available for 243 of the 280 samples. The first 10 principal components were used to adjust PRS. Imputation was performed using the caret k-Nearest Neighbours (KNN) imputation (knnImpute) (Kuhn, 2008).

4.2.3 Machine Learning Model for Classification

For classification purposes Generalised Linear Models were used. This was implemented with the R package caret (Kuhn and Johnson, 2013) using the glmnet package (Friedman et al., 2010) or the GLM R core package, in the case of the PRS only condition (where only a single predictor was used). Binary classification was performed using controls and first-episode psychosis.

Bootstrapping was used to generate 10,000 training datasets, with a sample size of 280 each. In each case, the samples not used for training were dedicated for testing. Caret was set up to use three values for the two tunable hyperparameters, alpha and lambda. This resulted in 9 hyper-parameter combinations. For each of these nine combinations, 10,000 were built on the training data and tested on the corresponding test set. This resulted in 90,000 models in total or 10,000 per hyperparameter combination. The average accuracy of all models for each of the nine parameter combinations was used to select the best settings.

Using the optimal hyperparameters, we created a final model on the full 280 samples. We saved test predictions of each of the 10,000 models and used the aggregate of these to estimate the accuracy and cohen's kappa of each model (referred to as kappa in the text). The kappa statistic compares observed accuracy with the expected accuracy of a random classifier, taking into account class imbalances. Kappa values can fall between -1 and 1, with a positive value suggesting better than random classification and 0 indicating performance equal to random predictions.

This process was repeated for each of the eight predictor combinations (Gx, Gx_Scz, PRS, Demo, PRS+Demo, Gx+PRS, Gx+Demo, Gx+Demo+PRS). Before creating bootstrap datasets for each predictor combination, the seed was set to 7, before training each model. The DoMC package in R to allow the use of 8 cores during model training was used.

4.2.4 Classification accuracy

Percentage of correct classification was calculated for each model, by pooling the results of all 10,000 bootstrap iteration predictions. Samples had a 37.5 % chance of being predicted in each iteration (Kuhn and Johnson, 2013), meaning each sample was predicted in approximately 3750 models (each bootstrap iteration built a new GLMNET model). Since assignment to the train and test data set is random at each iteration, there is slight variation in the number of predictions for each sample. A simple percentage was calculated for each sample, based on the instances each sample was assigned to their correct class and the exact number of models used to predict it. A value of 50% correct classifications for a given

sample thus means correct classification in 1875 models, while 100% would mean correct classification in all 3750 models.

4.3 Results

4.3.1 Polygenic Risk Score Imputation

Since Polygenic Risk Scores were only available for 243 out of 280 samples in the cohort, we decided to use imputation to avoid discarding samples for the machine learning models. We performed a simple KNN imputation, using only two predictors. This was based on a previous GAP study by Vassos et al. (2017), which demonstrated a strong effect between ethnicities, and identified differences in PRS between psychosis patients with a schizophrenia diagnosis, compared to other diagnoses or controls. The predictors chosen were, therefore, ethnicity and diagnosis. Ethnicity was stratified into Black, White, Asian and Other, while the diagnoses were stratified into Schizophrenia, Other Psychoses and Controls.

4.3.2 Performance of Classification Models

Eight GLMNET models were trained (see Table 4.1). The Estimated accuracy was highest for models using the entire gene expression data (Accuracy = 61%, Kappa = 0.22-0.23), with models 7 and 8 which incorporated Demographics performing marginally better as judged by their Kappa value (Table 4.1). PRS did not provide a noticeable benefit and was not selected by the GLMNET algorithm for any of the final models except models 3 and 5, which consisted of PRS alone and PRS in combination with demographics. The performance range across bootstrap iterations for all 8 models is visualised in Figure 4.1.

4.3.3 Classification accuracy across Bootstrap Iterations

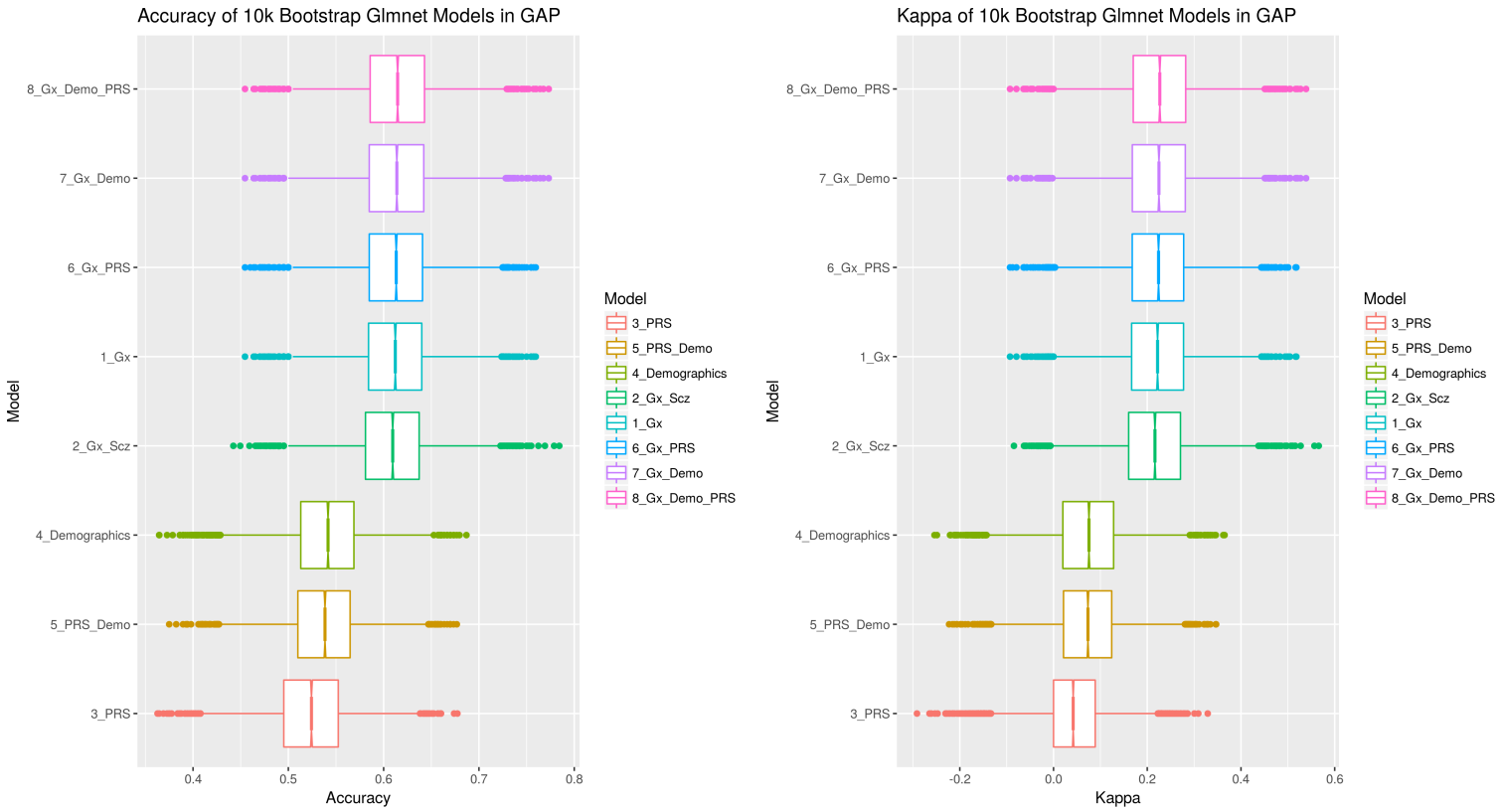
The percentage of correct classification was calculated based on approximately 3750 predictions, per sample, during bootstrapping. Density plots of samples were plotted based on this percentage (see Figure 4.2). The density plots were stratified by case-control status (Figure 4.2a) (FEP, Control), and by Diagnosis (Figure 4.2b) (SCZ, OP, Control). Models incorporating the full Gx data (1,2,6,7 and 8) had similar classification distributions. Controls were more accurately classified. When stratified by diagnosis, the Schizophrenia samples fell into a roughly bimodal distribution, with the larger peak at 90%, meaning these samples were misclassified in 10% of the bootstrap models (Figure 4.2b).

Table 4.1 List of models with estimated overall Accuracy and Kappa

Model	Features	Accuracy	Kappa
1	Gx	0.61	0.22
2	Gx_Scz	0.61	0.22
3	PRS	0.52	0.04
4	Demographics	0.54	0.07
5	PRS + Demo	0.54	0.07
6	Gx + PRS	0.61	0.22
7	Gx + Demo	0.61	0.23
8	Gx + Demo + PRS	0.61	0.23

Table of Machine Learning models used. Accuracy and Kappa for final model, incorporating all 280 samples, was estimated based on 10,000 bootstrap iteration. Models with Gx use the full gene expression data, Gx_Scz uses schizophrenia risk genes taken from Pirooznia et al. (2016). Demo refers to Demographics (Ethnicity, Age, Gender), while PRS is Polygenic Risk Score.

Across all bootstrap iterations Other Psychoses samples reached the highest density at 45% meaning they were classed as controls more regularly than Schizophrenia samples. This is illustrated further in Figure 4.3, where data from model 1 is visualised using boxplots stratified by Control and FEP (Figure 4.3a), or Diagnosis (Figure 4.3b). This shows that classification accuracy is higher for Schizophrenia patients, compared with patients diagnosed with Other Psychoses.

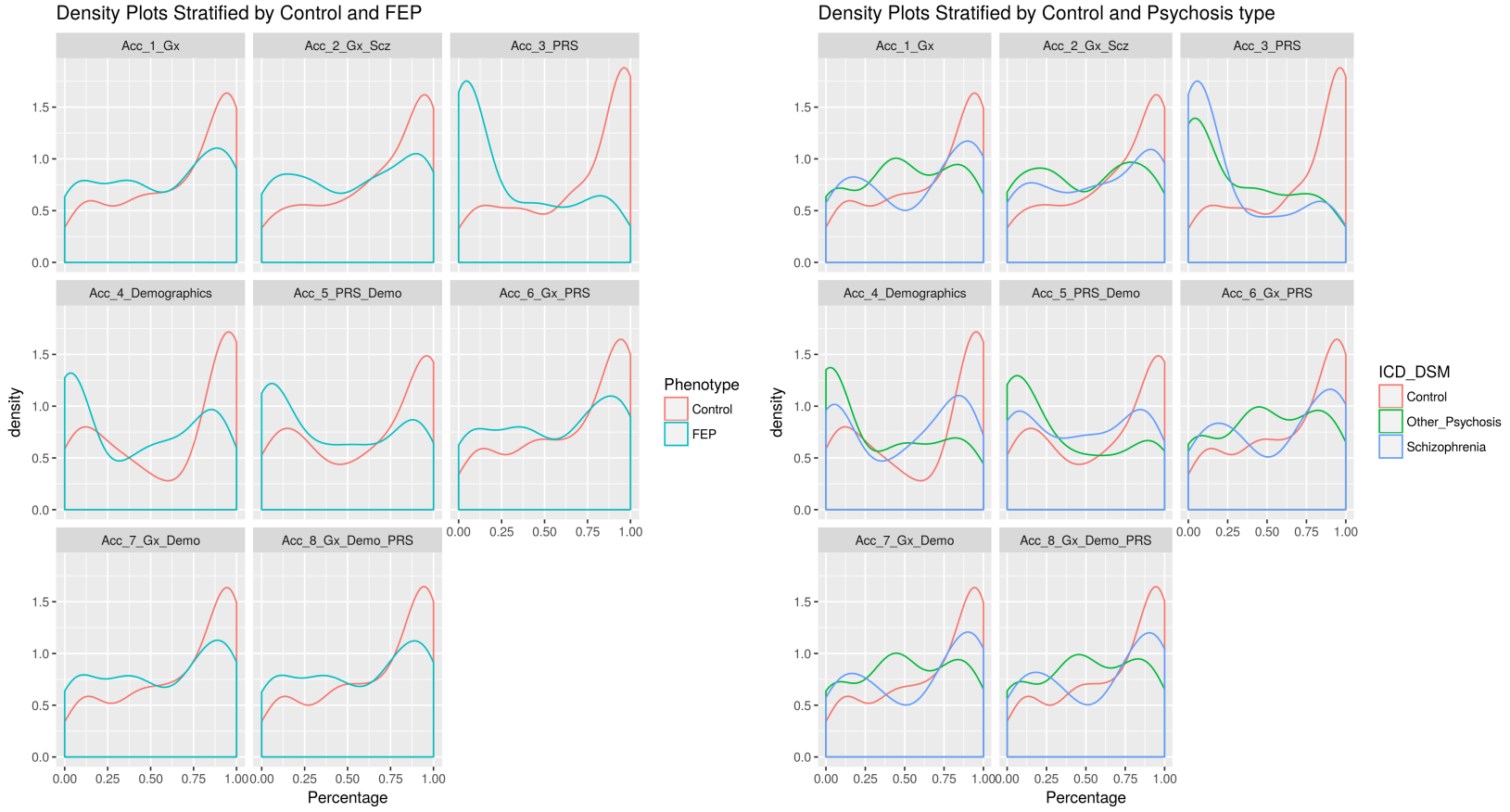


(a) Boxplot of Accuracy for GLMNET models

(b) Boxplot of Kappa for GLMNET models

Figure 4.1 Metrics for all 10k GLMNET models:

Metrics for all 10k GLMNET models: 8 models were constructed using GLMNET, using Gene Expression (Gx), Demographic (Demo), and Genetic data in the form of PRS. Models were built on variables from all three categories and all combinations (Gx, PRS, Demo, PRS+Demo, Gx+PRS, Gx+Demo, Gx+Demo+PRS). The 2_Gx_Scz model uses a hypothesis-driven subset of the overall gene expression data in Gx, which had prior Schizophrenia probability (Pirooznia et al., 2016). (a) Shows GLMNET models ordered by Accuracy. (b) Shows models ordered by Kappa.

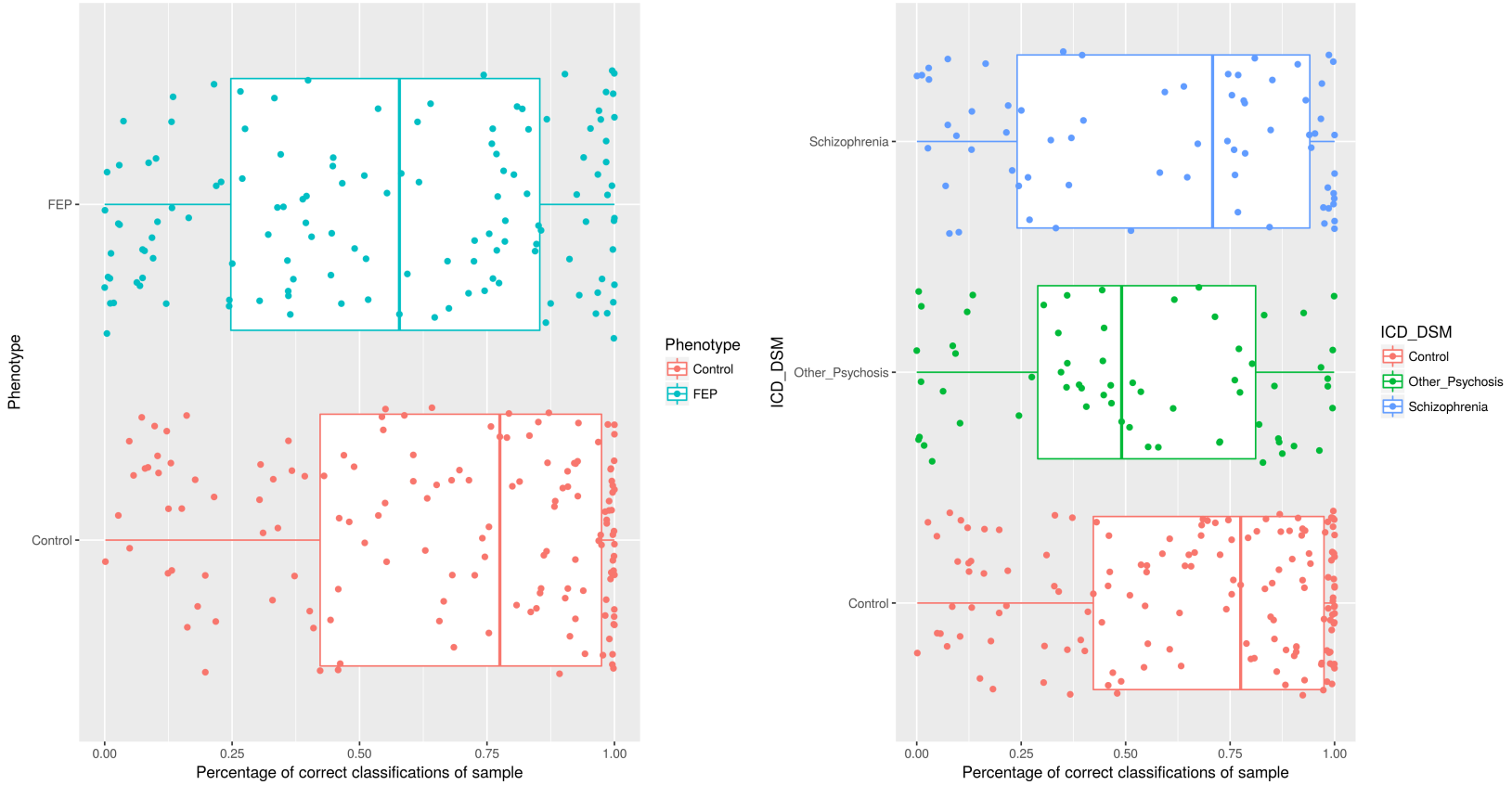


(a) Stratified by Control and FEP

(b) Stratified by Diagnosis

Figure 4.2 Density Plots for all 8 Glmnet models:

Density plots of all 8 GLMNET models. (a) Shows GLMNET stratified by FEP and Control. (b) Shows GLMNET stratified by Schizophrenia, Other psychoses and control. The y-axis shows density of samples, while the x-axis shows the percentage of correct classifications for each sample. Each sample was predicted, on average using 3780 different GLMNET models. The results were used to calculate percentage of correct classifications.



(a) Stratified by Control and FEP:

(b) Stratified by Diagnosis

Figure 4.3 Boxplots of classification accuracy in Gene Expression (Model 1):

Box plots for GLMNET Gene expression model (model 1). Samples were on average predicted in 3750 models (bootstrap iterations). The percentage of correct classifications for samples across all models was used to plot samples (50% meaning correct classification in 1875 iterations), with each point representing one sample. (a) shows samples stratified by FEP and Control. (b) shows samples stratified by Schizophrenia, Other Psychoses and Control. The y-axis variation for individual points within a group, is unrelated to the data, and is purely a visual aid to clarify the distribution of points across the x-axis.

4.3.4 Psychosis severity correlated with classification accuracy

Full Positive and Negative Syndrome Scale scores were available for 49 out of 68 SCZ patients, and 47 out of 63 OP patients. The two populations were split into quantiles based on Classification accuracy from bootstrap iterations in model 1. The quantile splits were made to create 4 groups of equal size, and fell at 0%-25%, 25%- 58%, 58%-85% and 85%-100%. Figure 4.5 shows full PANSS and subscales plotted against accuracy quantiles.

Median PANSS was lowest for Schizophrenia samples in the first quantile (0%-25% correct classification), for all subscales with the exception of the Negative symptom subscale (see Figure 4.5c), where the highest median score was observed. Similarly the highest median PANSS score in the schizophrenia samples was observed in the last quantile containing the most accurately classified samples across all iterations (85% - 100% correct classification). This was, again, on all but the negative symptom subscale. Correlations between PANSS and classification accuracy in schizophrenia was 0.3 for positive symptoms (p-value = 0.03), 0.25 for overall PANSS (p-value = 0.08), 0.22 for general psychopathology (p-value = 0.11) and 0.01 for negative symptoms (p-value = 0.96).

For schizophrenia samples there was therefore an overall trend of higher scores on the Positive subscale that was significant. The trend was also observed for the Psychopathology subscales for samples that were more accurately classified, however this was not significant. This increase in median PANSS was observed for schizophrenia but not for other psychoses, where the highest correlation of 0.1 was seen with negative symptoms, but this failed to reached significance (p-value = 0.49). Schizophrenia patients with high PANSS scores on the positive subscale were more accurately classified, while that was not true for other psychoses.

4.4 Discussion

4.4.1 Classification Accuracy was highest for Schizophrenia linked Psychosis

The predictive power of the models was lower than previous studies, with none reaching an estimated accuracy above 61%. In contrast, a recent study by Hess et al. (2016) achieved an estimated accuracy above 90% and another study by Lee et al. (2012) had similar results in blood. However, these result in came from more homogeneous dataset consisting of chronic schizophrenia patients. GAP is a multi-ethnic cohort, and only about half the patients qualify for a Schizophrenia diagnosis according to the Operational Criteria Checklist for Psychotic Illness and Affective Illness (OPCRIT) system (Rucker et al., 2011). The discrepancy in predictive power could be due to confounding with medication in other studies or different

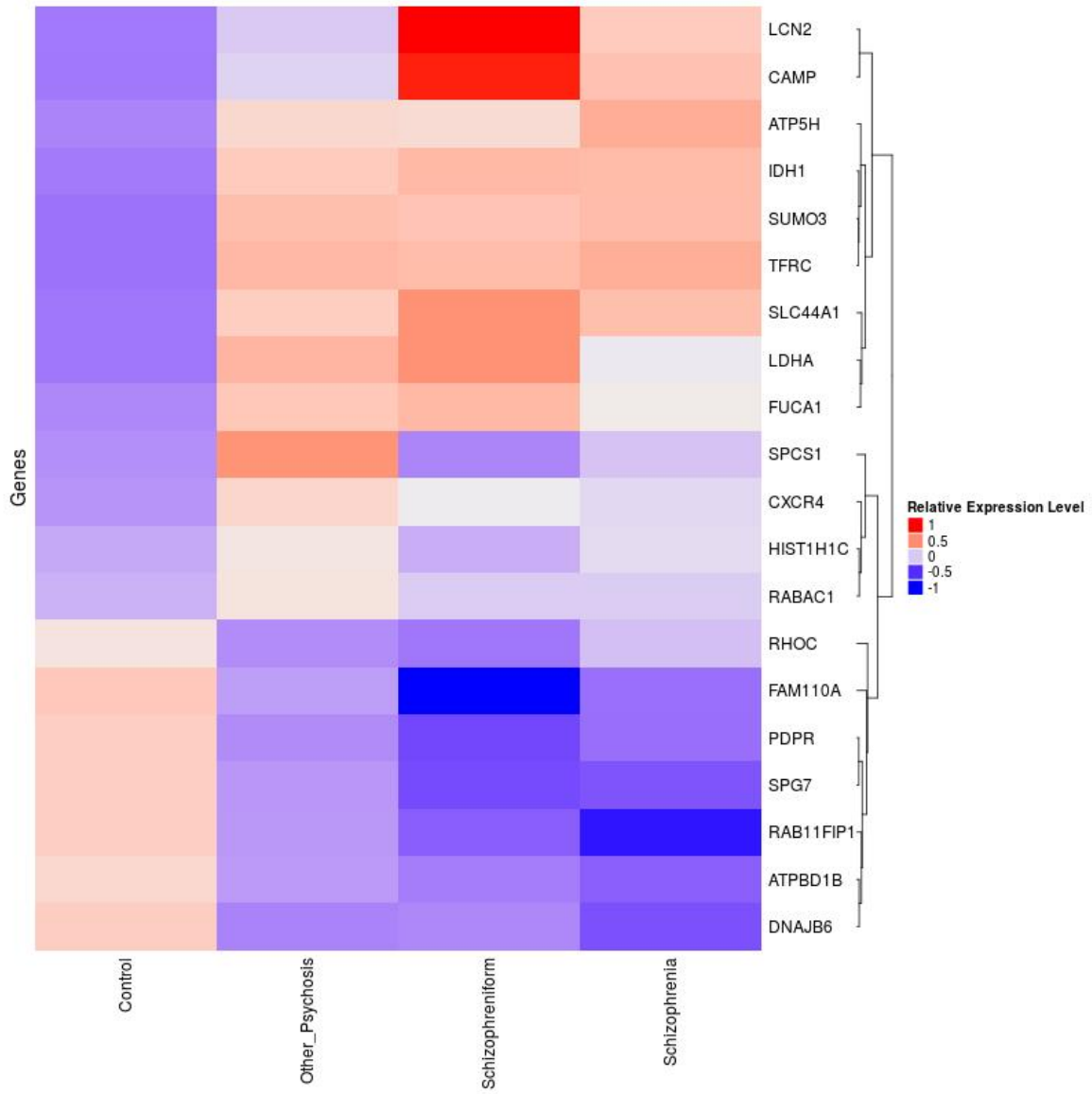
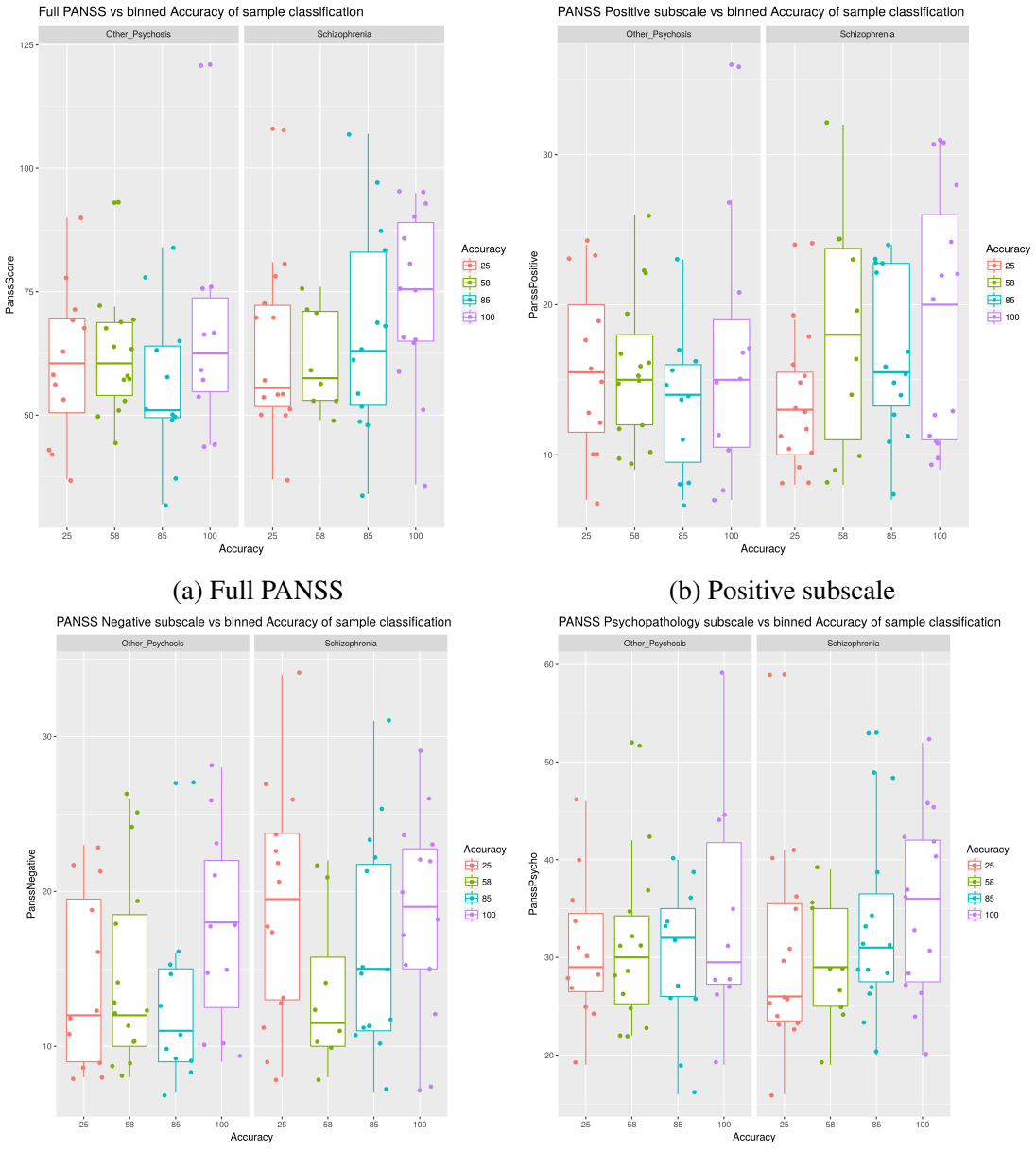


Figure 4.4 Heatmap of top 20 features of Gene Expression Model 1:

The top 20 predictive genes from Model 1 (all gene expression) are plotted. Gene expression values were centred and scaled for each feature, and averaged across each subcategory. Samples are stratified by Control, Other Psychosis and Schizophrenia as before, with the exception that the schizophrenia category was split into Schizophrenia and Schizophreniform disorder according to the DSMIV diagnosis. Rows are clustered using complete-linkage clustering, with the dendrogram indicating distance.



(a) Full PANSS

(b) Positive subscale

(c) Negative subscale

(d) Psychopathology subscale

Figure 4.5 Boxplots of PANSS plotted against binned levels of classification accuracy:

Classification accuracy was split into quantiles, resulting in accuracy categories 0%-25%, 25-58%, 58%-85% and 85%-100%. Patients with Schizophrenia (n = 49) are on the right and Other psychoses (n = 47) are on the left. (a) Shows overall PANSS (b) Shows the Positive PANSS subscale (c) Shows the Negative PANSS subscale (d) Shows the general Psychopathology PANSS subscale.

computational approaches. An interesting possibility is that Schizophrenia has a more consistent phenotype, or that the blood signature in other psychosis is more heterogeneous.

The highest accuracy was achieved by model 8 which uses Gene Expression, Demographics and PRS. However, this constituted a minimal improvement over any model that included Gene expression data, with accuracy increases far below 1%, meaning that neither PRS or demographic variables provided a benefit over Gx alone.

When examining the density plots of models using just gene expression data (Model 1), the highest density of control samples was observed at approximately 95% accurate classification across all bootstrap iterations, with a sharp drop as accuracy decreases. First episode psychosis cases, on the other hand, have a largely flat distribution, with a small peak at 95% (see Figure 4.2b), suggesting a highly heterogeneous gene expression signature that is distinct from controls for only a subset of samples. Interestingly when the data is split into diagnostic categories, the schizophrenia distribution forms two peaks, one at 90%, and a smaller one at 10%. This suggests that the GLMNET can identify a distinct and accurate gene expression signature for a small majority of schizophrenia samples, while a large minority is consistently misclassified.

Samples with diagnoses of other Psychoses on the other hand peak at 50%, meaning most Other Psychosis samples are essentially classified at random. This indicates a heterogeneity between Schizophrenia and Other Psychosis within their gene expression signature.

4.4.2 Polygenic Risk Score provided no improvement in predictive power

The polygenic risk score model achieved an estimated accuracy of 0.52 which provided little benefit in classification. None of the models using gene expression data selected PRS as a predictive variable. While PRS has been shown to have some power to distinguish schizophrenia samples from controls, it has limited power in populations of African descent (Lu et al., 2014) due to the allele frequencies being more diverse in those populations. Since this cohort is multi-ethnic and first-episode, with only half the patients receiving a schizophrenia diagnosis, any predictive power is likely significantly reduced. As the PGC (Ripke et al., 2014) increases its sample size PRS predictive power is presumed to increase. This may lead to better predictions in the future when combined with expression data.

4.4.3 Schizophrenia and Schizophreniform disorder comparison

Since our Schizophrenia category uses the OPCRIT system which synthesises ICD-10 and DSM-IV criteria, approximately half the samples categories as Schizophrenic suffer from Schizophreniform disorder under the DSM-IV criteria. Schizophreniform disorder is a less

severe form of Schizophrenia, with a better outcome. Diagnosis occurs when patients have not met the full diagnostic criteria for schizophrenia. This usually means symptoms have developed no more than six months before diagnosis.

As such, I hypothesised that misclassified samples in the Schizophrenia group may correlate with schizophreniform disorder under the DSM-IV, under the assumption that gene expression signature may not have been as prominent in these samples. No evidence for this was found, in fact, the trend was slightly reversed, with the median schizophreniform patient being more accurately predicted. Since outcome for schizophreniform disorder tends to be better, despite the symptom overlap, this may suggest that schizophreniform is more strongly influenced by environmental factors in the first episode when compared to long-term symptoms in schizophrenia, and a gene expression signature may be more prominent in these individuals. Reduction in stressful life events and provision of care could conceivably be consistent with higher rates of recovery, while genetic vulnerability and structural changes in the brain would presumably play a more significant role in schizophrenia and lead to the worse long-term outcomes. This is speculative however and would need to be investigated in a larger sample size.

4.4.4 Genes Important for Model Performance are linked to Immune System

The gene expression heatmap of the top 20 predictors in model 1 (Figure 4.4), showed slight expression changes between controls and other psychosis. The strongest signals, however, were between schizophreniform disorder, and to a lesser extent schizophrenia and controls. In model 1, which used the full gene expression data, we note Lipocalin 2 (LCN2) and Cathelicidin Antimicrobial Peptide (CAMP) upregulation, especially in Schizophreniform disorder. These probes cluster most closely with SUMO3, ATP5H, IDH1, TFRC, SLC44A1, LDHA and FUCA1 which altogether forms a signal of upregulation in all three groups compared to controls.

CAMP is a member of a class of proteins involved in innate immunity, primarily in macrophages. CAMP expression is also mediated by vitamin D, which has been studied extensively in regard to schizophrenia, depression, the immune system and sleep (Anderson and Maes, 2012; Aranow, 2011; Brown and Roffman, 2014; McCarty et al., 2014; Moylan et al., 2014; Penckofer et al., 2010). It also is correlated with Defensin expression, which forms the most upregulated proteins in our study (see Chapter 3). They are markers of inflammation, stress and upregulation of the innate immune system.

LCN2 (also known as NGAL) and Transferrin receptor 1 (TFRC) have also been shown to play a role in the innate immune system (Ganz, 2009; Goetz et al., 2002), notably in an IL-6 mediated sequestering of iron, which impedes the survival of many pathogens (Parrow et al., 2013). Interestingly, intracellular pathogens also make use of this mechanism to increase iron available within the cell (Parrow et al., 2013). Both LCN2 and TFRC are up-regulated (logfc 0.58 and 0.17 respectively) in the FEP patients (Chapter 3), which suggests iron uptake from the serum. Chronic stress is known to significantly reduce serum iron levels (Wei et al., 2008). Iron deficiency is more common in psychiatric patients, with 35% of psychosis patients being affected according to one study (Korkmaz et al., 2015). In addition increased accumulation of iron in the basal ganglia (in which TFRC plays a role) is linked to a range of neurodegenerative disorders such as Huntington's and Alzheimer's (Wong and Duce, 2014).

SUMO3 much like ubiquitin can act as a protein-based post-translational modification and plays a vital role in numerous pathways, including NF-κB signalling (Frank et al., 2013). The role of SUMO is complex and multifaceted, but one study found SUMO3 knock-downs increased NF-κB signalling following TNF-α stimulation (Frank et al., 2013). SUMO3 is the most significant differentially expressed (see chapter 3), although not the most up-regulated. A stress hypothesis, via the HPA-axis would suggest NF-κB up-regulation at the protein level, and the differential expression results in chapter 3 did show enrichment of pro-inflammatory pathways.

ATP5H and IDH1 are involved in mitochondrial function. Studies have also identified SUMO3 and NF-κB upregulation following glucose and oxygen deprivation (Sirabella et al., 2009; Yang et al., 2012). Since mitochondria play a core role in oxidative stress pathways, and these pathways are enriched in psychosis expression studies (Hess et al., 2016), the above-mentioned genes may be plausible biomarkers linked to schizophrenia pathology.

The probes selected by the model, in combination with differential expression evidence previously laid out, point to a signal rooted in innate immunity, glucose metabolism, oxygen deprivation and iron depletion. The signal is consistent with a pattern found in Stress, which can be mediated by a variety of factors including psychological ones, chronic sleep deprivation, infections, oxygen and glucose deprivation, environmental exposure and genetic vulnerabilities. Of interest are differences between Schizophreniform disorder, Schizophrenia and Other Psychosis in this context. The strongest signal was found for LCN2 and CAMP between controls and schizophreniform disorder. These genes were also upregulated in schizophrenia, but to a lesser extent, while they showed a slight downregulation in other psychosis. Both LCN2 and CAMP may be useful as biomarkers for schizophreniform disorder and identify patients who have a higher chance of recovery. This would need to be replicated however.

4.4.5 Classification Accuracy was associated with Positive Symptom Severity

The four PANSS groups were correlated with the classification accuracy of Schizophrenia and other psychosis samples separately. Classification accuracy's were taken from model 1. The results indicated a significant correlation between positive symptoms and accurate classification of schizophrenia samples, but not for other psychoses. None of the other PANSS subscales achieved significance at a p-value threshold of 0.05. This suggests that the most prominent blood signature that the classifier is using is related to positive symptoms, specifically in schizophrenia. This mirrors results from chapter 3 indicating a correlation of positive symptoms with innate immunity.

The overall low accuracy of the binary FEP vs Control classifier might thus be related to the heterogeneity between schizophrenia and other psychoses. What these results show however is that the GLMNET classifiers preferentially identify schizophrenia samples with more severe symptoms, which indicates a common underlying gene expression signature. These result may be an artefact of confounding with medication or drugs, but since samples with known drug induced psychosis were excluded and little evidence of a medication effect was identified in chapter 3 this seems unlikely.

The probes selected by the model are involved in innate immunity and the stress responses, and are up or down regulated in a manner that is consistent with the schizophrenia literature. While immune deregulation also plays a role in disorders like depression and bipolar disorder, the magnitude and precise molecular signature may be different. Alternatively it is possible that less prominent features in the model provide the necessary specificity for the schizophrenia signature.

While the DSM and ICD categorisations have been criticised for their lack of biological support, this may provide evidence that schizophrenia is at least more coherent than general non-pharmaceutically induced psychoses, in terms of the gene expression signature.

4.5 Conclusion and Future Directions

Overall our results achieve much lower accuracy than has been previously reported for blood gene expression models. This is likely due to greater heterogeneity in the sample, both in terms of demographics and diagnosis.

I found no evidence that genetic information in the form of PRS significantly increased model performance. Again, this may be due to the high percentage of patients of African ancestry. Alternatively, genetic risk may already be captured by gene expression. The

features that were selected are consistent with a stress model of psychosis, based on HPA axis arousal. A positive and significant correlation between classification accuracy and positive symptoms in schizophrenia samples, but not in other FEP samples was found. As it stands, this supports the idea that more severe positive symptoms are related to stress signals and are primarily responsible for distinguishing cases from controls. In future work, it may be useful to include adverse and traumatic life events in the models, which was not possible in this case. It would also be valuable to test the hypothesis that the gene expression signature in Scz is more coherent than in other psychoses, by building models in other psychiatric datasets (especially affective disorders) and apply them to this cohort.

While these results have limited clinical value, further dissection of gene expression signatures in Psychosis may provide valuable insights into the underlying nature of a patients psychosis. The current evidence shows consistent but non-specific disruption of pathways across multiple psychiatric disorders. Using gene expression data from chronic stress, autoimmunity, pathogenic infections, sleep deprivation, drug use, and other stressors, it might be possible to identify biomarkers that indicate a more subtle disruption in gene expression. This could theoretically allow gene expression signatures to determine if pathway disruptions are related to psychological stress, infections or genetics, which might ultimately help guide treatment decisions.

Chapter 5

Predictive Modelling using the Dejong data

5.1 Introduction

First Episode Psychosis has serious clinical implications and presents a challenge for physicians in terms of selecting an appropriate treatment from a range of pharmacological and therapy options. Psychotic symptoms range from hallucinations and delusions to thought disorder and catatonia and have a wide variety of contributing factors including genetic, neurological, environmental and substance abuse. Complicating this is that subdivision of psychotic disorders in either the International Classification of Diseases 10 (ICD-10), or the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), are still deeply rooted in Kraepelin notions, stemming from the early 20th century. While the definitions of Psychotic disorders such as Schizophrenia, Bipolar Disorder and others, have evolved, current diagnostic techniques are still quite subjective.

Outside of substance-induced psychosis, perhaps the clearest examples of progress in classifying psychosis, is the recent realisation that a small percentage of apparent schizophrenia cases are suffering from NMDAR antibody Encephalitis (Zandi et al., 2011), a disease characterised by antibodies with glutamate receptor reactivity. This represents a significant step towards a more individualised and targeted treatment approach.

Genetic studies have become more sophisticated in identifying genes linked to schizophrenia risk (Ripke et al., 2014). However these markers as of yet are of limited clinical use, and physicians often have little option, but to cycle through a series of interventions, by trial and error, until one is found that is tolerable for the patient. Due to the heterogeneous nature of psychosis, gene expression studies can potentially provide a resource of biomarkers that

sit at the intersection of environmental and biological factors. Ideally, this would help in identifying diagnostic subdivisions and help with patient care in the clinic, by moving beyond symptom-based diagnostic and treatment approaches. While attempts have been made to use blood gene expression data to distinguish psychosis patients from controls in the past (Lee et al., 2012), the literature is lacking in replication and often suffers from a lack of samples. A notable exception to this is the schizophrenia mega-analysis by Hess et al. (2016) which pooled available transcriptomic data from 8 studies (including de Jong et al. (2012)) and used SVM and Random forest approaches to build classifiers designed to distinguish case-control status. Their study design trained on data from Illumina chips and used Affymetrix data for validation. They managed to achieve an AUC of above 0.9 in the training data and above 0.7 in the test data.

Due to the heterogeneous nature of the GAP data available to us, in addition to our previous results showing poor predictive performance within GAP and for schizophrenia datasets (chapter 4), we wanted to investigate if predictive models built on schizophrenia would perform better. Our previous results indicated differences between schizophrenia and psychosis samples with other diagnoses, but since the models were built on FEP as a single category, and without test data for final models, it is necessary to create additional classifiers on external data. This is important to verify the assumption that schizophrenia is a more homogeneous group and to provide insight into the molecular signature of other FEP cases. For this purpose, a previously published de Jong et al. (2012), and well define chronic schizophrenia cohort was identified, consisting of 239 northern European individuals. A 202 sample subset was chosen to build a Schizophrenia classifier and to test it on the GAP data.

We hypothesised that the low predictive power of machine learning models trained on first episode psychosis is related to the heterogeneous nature of FEP. To support this hypothesis, we aimed to build classification model on DeJong data and test the performance on the GAP dataset. We hypothesised that a robust model would more accurately classify first episode psychosis patients, who are of European ancestry, have higher symptom severity, ultimately got a diagnosis of schizophrenia, and took anti-psychotic medication. Finally, we hypothesised that individuals with disorders thought to be closely related to schizophrenia, based on genetic studies, heritability, and overall clinical presentation, would be classified more accurately. Roughly speaking, the assumption was that accuracy would decrease progressively across the spectrum from schizoaffective disorder, delusional disorders, bipolar disorder to depression.

5.1.1 Aims

In order to test this hypothesis we had 4 core aims.

1. The first was to train a series of predictive models, including ensemble models using the dejong data, and to provide evidence that these models are robust and accurate using internal validation in the form of training sets and cross-validation.

2. The second aim of this project was to validate the models in GAP and report the performance.

3. The third aim was to investigate differences in classification confidence and probability for GAP patients who later got a diagnosis of schizophrenia and those who were diagnosed with other psychoses. We aimed to investigate any signal further by looking for consistency across all trained machine learning models.

4. The fourth aim was to identify potential clinical or biological confounding factors that may mask or generate a signal. Of special concern here was anti-psychotic medication and ethnicity.

5.2 Methods

5.2.1 Gene Expression Datasets

For training Machine Learning models in this chapter we used the DeJong Chronic Schizophrenia cohort (referred to as Dejong after this point) as described in de Jong et al. (2012). The 239 samples used in the study came from 2 platforms, 202 from Illumina H-12 and 37 from Illumina H-8. To simplify the process the 37 samples from the Illumina H-8 chip were not incorporated in this analysis. The full data for the other samples is freely available at the ArrayExpress Archive (https://www.ebi.ac.uk/arrayexpress) with the identifier E-GEOD-38484. The E-GEOD-38484 gene expression data was obtained using the ArrayExpress R package (Rustici et al., 2013). The same subset of the GAP data was used as in chapters 3 and 4 namely 280 samples (149 controls, 131 FEP).

5.2.2 Processing

All datasets were subset to probes that overlap with Gene Symbols identified in the GAP dataset as described in the preprocessing chapter. This was achieved in the DeJong Chronic Schizophrenia cohort (Dejong) dataset by using illuminaHumanv3.db package (Dunning et al., 2015) to translate Illumina ProbeIDs to nuIDs. Probes in the GAP dataset that did not have any direct nuID matches in the DeJong dataset were examined again using nuIDs, which correspond to the same Gene Symbol. All remaining probes were dropped in all datasets.

Probes with LOC or HS. Prefixes were also removed. We used Cellmix (Gaujoux and Seoighe, 2013) and Surrogate Variable Analysis Leek et al. (2012) to identify any potential additional confounding factors. All datasets were centred and scaled before any further analysis.

5.2.3 Machine Learning

The machine learning process involved multiple steps. External data was randomly split into train and test data 10 times. The training data sets were reduced to relevant features, by predefined cut-offs, and random forest based recursive feature elimination (RFE). The remaining features were used to build models with 6 machine learning algorithms of different families. The models were pooled for each of the 10 splits and an ensemble was created using stochastic gradient boosting (GBM). Models were validated in test data, and finally used to predict GAP data. This process is described in detail below and in figure 5.1.

Train and Test data

Ten sets of Train and Test data were randomly generated. This was done using the "createDataPartition" function from the caret package Kuhn (2008), by selecting 80% of samples for the training data and using the remaining 20% as test data for that set. The ratio of Schizophrenia samples to Controls was kept constant in all datasets. Down-sampling was performed at later stages during the model building process.

Feature Selection

Feature selection was performed on each of the 10 Training data sets separately. This was done in 3 steps. The first step was the removal of highly correlated features; these were defined as features with a Pearson correlation coefficient above 0.75. The second step was the removal of features with a variance across samples below 0.25. The third step was an implementation of RFE incorporating re-sampling from the caret package Kuhn (2008), by using the rfe function, with slight modification (Algorithm 1). The training data was centred and scaled, and the RFE algorithm was run using 30 bootstrap iterations, with downsampling. A predefined list of 30 feature sizes (750, 725, 700 ... 25) was passed on to the function. a Random Forest algorithm (with OOB settings), was used to build models. Features were ranked by importance internal to the caret package. The top-performing feature size was selected, and in situations where the number of selected features exceeded the number of

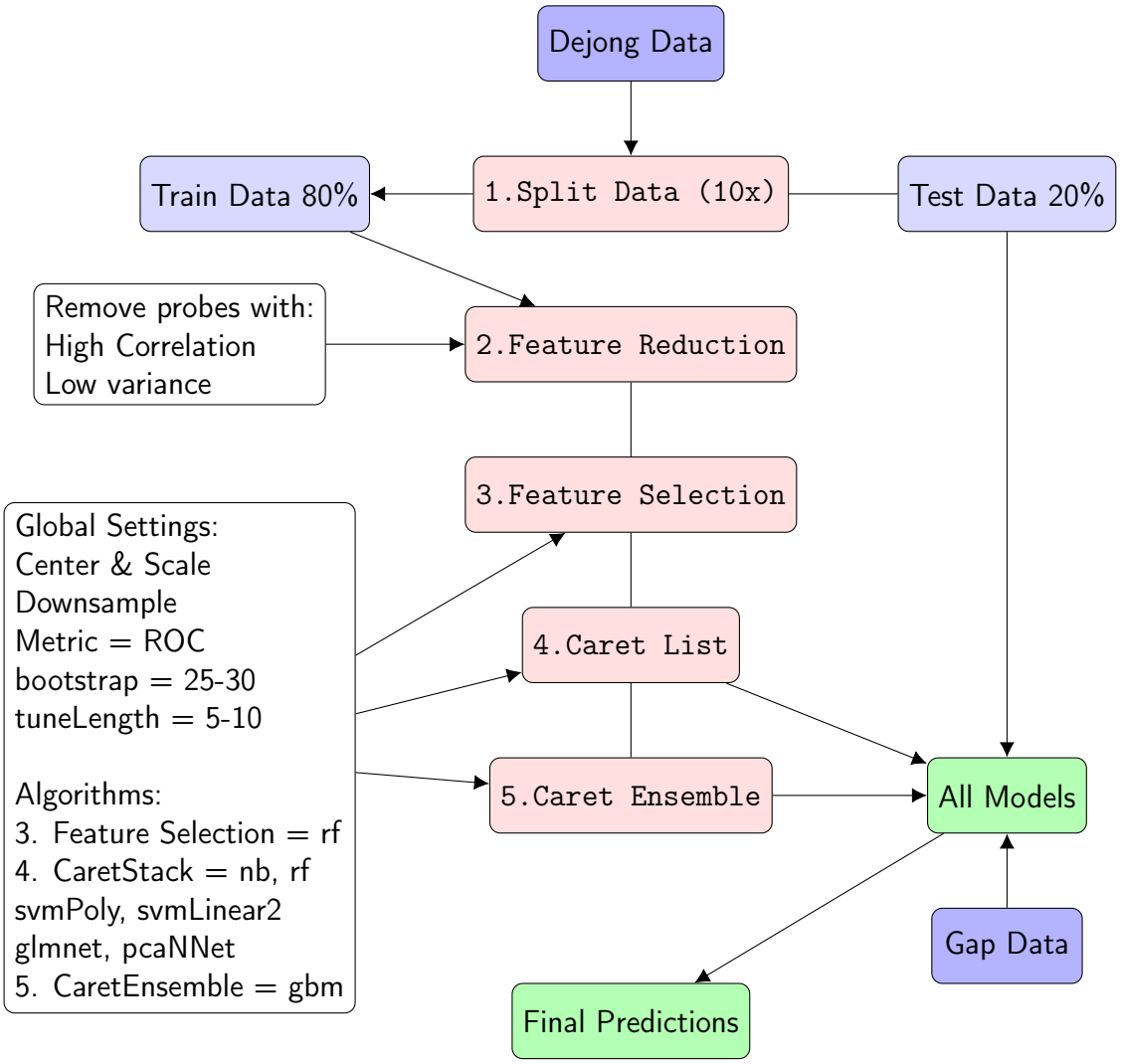


Figure 5.1 Machine Learning Ensemble Flowchart

Flowchart of steps in Machine Learning Ensemble process. *1.* cleaned Chronic Schizophrenia data de Jong et al. (2012) was randomly split 10 times into training and testing data. *2.* Feature reduction was performed on each of the 10 training datasets, which included removal of highly correlated and low variance probes. *3.* Recursive feature selection using Random Forest, with 30 bootstrap iterations, was implemented. *4.* Six machine learning algorithms (nb, rf, smvPoly, svmLinear2, glmnet, pcaNNet) were implemented with 25 bootstrap iterations on each of the 10 training datasets. All bootstrap iterations used the same sample indices. *5.* Caret Ensemble was implemented using Stochastic Gradient Boosting. An Ensemble was built on the 6 models from step 4, for each training set, using 25 bootstrap iterations. All classification models from steps 4 and 5 were used to predict the GAP data and corresponding test data, set aside in step 1.

samples, only the N most important features were taken forward, where N is the number of samples in the training set.

Data: 10 Dejong Training Data sets
Result: 10 Training Data sets with reduced features
for *each Training Data set* **do**
  for *each resampling iteration* **do**
    Generate train and test data using bootstrapping;
    Train model on train data with Random Forest using all P predictors;
    Calculate model performance;
    Calculate variable importance;
    for *Each subset size Si (750, 725,700 ... 25)* **do**
      Keep the Si most important variables ;
      Train model on train data with Random Forest using all Si predictors;
      Calculate model performance;
      Calculate variable importance;
    end
  end
  Calculate performance over Si on test data;
  Determine top performance model;
  Rank predictors of model;
  if *P predictors > N Sample Size* **then**
    Select N top predictors;
  end
end
Save Predictor lists for downstream model building;

Algorithm 1: Recursive Feature Selection

Training of Models

The Training data were subset to the features selected previously and was scaled and centred again. Machine learning models were trained following this using the caretList function, from the caretEnsemble package (developer version, 24th may 2017, tinyurl.com/cEnsemble) in R. The structure of the process can be seen in Algorithm 2. In short, six machine learning algorithms were selected (see Table 5.1) and individually passed from caretList to the train function from the caret package. Random hyperparameter search was implemented via caret Kuhn (2008) , using the tuneLength command. The search was performed across ten values per hyperparameters, except in Random Forest and Naive Bayes, where the search was across

30 and five values respectively. For Random Forest only the mTry parameter was searched, and this was manually done across all values from 1 to 30. In the case of Naive Bayes, the tunable hyperparameter space was fully explored.

All models used bootstrapping for purposes. Twenty-Five Intermediate Train and Test data sets were generated and used in the generation of all models and for all hyperparameter combinations. The optimal model was selected using the receiver operating characteristic (ROC), and the final model was trained on the initial training data using the same settings.

```
   Data: Dejong Training Data
   Result: Machine Learning Models
 1 Generate 25 bootstrapping train and test data indices;
 2 for each machine learning algorithm a do
 3     for each hyperparameter combination h in a do
 4         Generate train and test data using bootstrap index i;
 5         for each resampling iteration i do
 6             Train model a on train data i using hyperparameters h ;
 7             Calculate model performance in test data i;
 8         end
 9         Find top performing hyperparameters;
10     end
11     Train model a on full input training data using final hyperparameters;
12     Add model a to list of final models;
13 end
14 Save list of final models
```

Algorithm 2: Machine Learning using caretList

Training of Ensemble Model

The ensemble model was trained using the 6 models generated previously. This was achieved using the caretStack function from the caretEnsemble package. GBM was used to produce an ensemble model from the 6 constituent models. This was implemented via the gbm package. AUC derived from ROC was used as the evaluation metric, we used bootstrapping (n = 25) for resampling, and a random parameter search (tuneLength = 10) to identify the optimal hyper-parameters.

Table 5.1 List of Machine Learning Algorithms used

Algorithm	Algorithm Full Name	R Library	Tunelength
svmLinear2	SVM with Linear kernel	e1071	10
svmPoly	SVM with Polynomial kernel	kernlab	10
glmnet	GLM with Lasso or Elastic-Net	glmnet	10
pcaNNet	ANNs with Principle Component Step	nnet	10
nb	Naive Bayes	klaR	5
rf	Random Forrest	randomForest	30
gbm	Stochastic Gradient Boosting	gbm	10

Table of Machine Learning algorithms used and the names of corresponding R libraries called by the caret package. Tunelength refers to the number of values used for each tunable hyperparameter. SVM stands for Support Vector Machines, GLM stands for Generalised Linear Models, ANNs stands for Artifical Neural Networks. Stochastic Gradient Boosting was used for the ensemble model.

5.2.4 Testing performance in external data

To test performance GAP and all other external datasets were scaled and centred individually. Following this, the final models for all 7 algorithms (6 caretStack, 1 ensemble) were used to predict class probabilities for all samples within the current dataset. The predict function from the caret package was used for this.

5.2.5 Testing of variables associated with classification accuracy

Following prediction, GAP FEP cases were split into Schizophrenia and Other_Psychosis based on ICD-10 and DSM-IV diagnostic criteria. If either criterion labelled the patient as schizophrenic, they were put into the schizophrenia group. Otherwise, they were classed as Other_Psychosis. Subsets were evaluated separately for predictive performance. Additional tests were performed to identify variables correlated with predictive confidence for the ensemble model predictions in GAP data. The variables tested were Ethnicity, Gender, Tobacco use, Medication, PANSS and Diagnosis (ICD-10 and DSMIV). Variables were tested independently, and samples were excluded in each instance based on the available clinical information. One way ANOVA was used, in combination with the Tukey's range test.

5.3 Results

5.3.1 Demographics

The demographic data available for participants in the deJong study was limited. We had access to information on Gender and Age (see Table 5.2). There was a statistical difference in gender between Schizophrenia patients and controls. The GAP cohort was identical to previous chapters, FEP patients were split up into Other Psychosis and Schizophrenia for some analysis, as in chapter 4. Demographic data for all 3 groups can be found in Chapter 2, Table 2.1.

Table 5.2 Dejong Chronic Schizophrenia Demographics

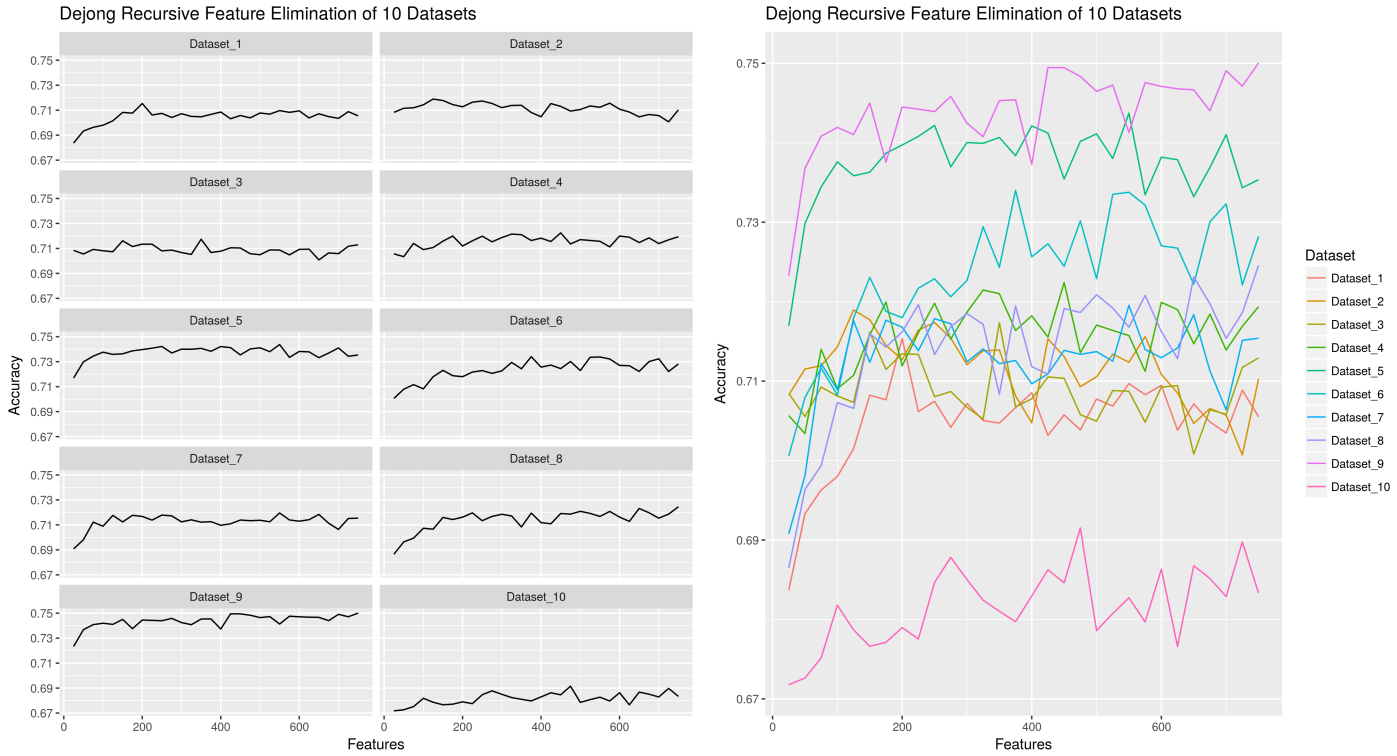
	Control	Schizophrenia (SCZ)	p-value
n	96	106	
Gender = male (%)	42 (43.8)	76 (71.7)	<0.001
Age (mean (sd))	39.31 (14.19)	39.58 (10.74)	0.877

Table of demographics for the subset of data used from de Jong et al. (2012). Data is split by control and schizophrenia. Only information for gender and age was available. A significant difference was found for gender between groups. P-values were calculated using the chi-square test for Gender, and t-test for age.

5.3.2 Machine Leaning classifiers built in Chronic Schizophrenia

After comparing GAP and Dejong data, 3918 genes were found to be expressed in both. The machine learning pipeline was implemented as described in figure 5.1. First, a cross validation algorithm was used to generate 10 sets of training, and 10 corresponding sets of test data. At each iteration 80% of samples (n = 162) were assigned to the training data, and the remaining 20% (n = 40) to the test data. Seven machine learning models were trained for each data split, six on the training data and one ensemble model built on six initial models (See Table 5.1), resulting in a total of 70 models.

Estimated accuracy during the feature selection approach varied between 67.5% and 75% (Figure 5.2). The biggest increase in accuracy was seen by increasing features from 25 and 200 features. Datasets 5, 6 and 9 showed the highest performance above 73% while dataset 10 showed the lowest at 67%. Datasets 2 and 3 were most consistent in estimated accuracy across all features at 70% - 72% accuracy. Dataset 2 is the only one that showed the highest performance with fewer probes than samples, selecting just 125.



(a) RFE Facet

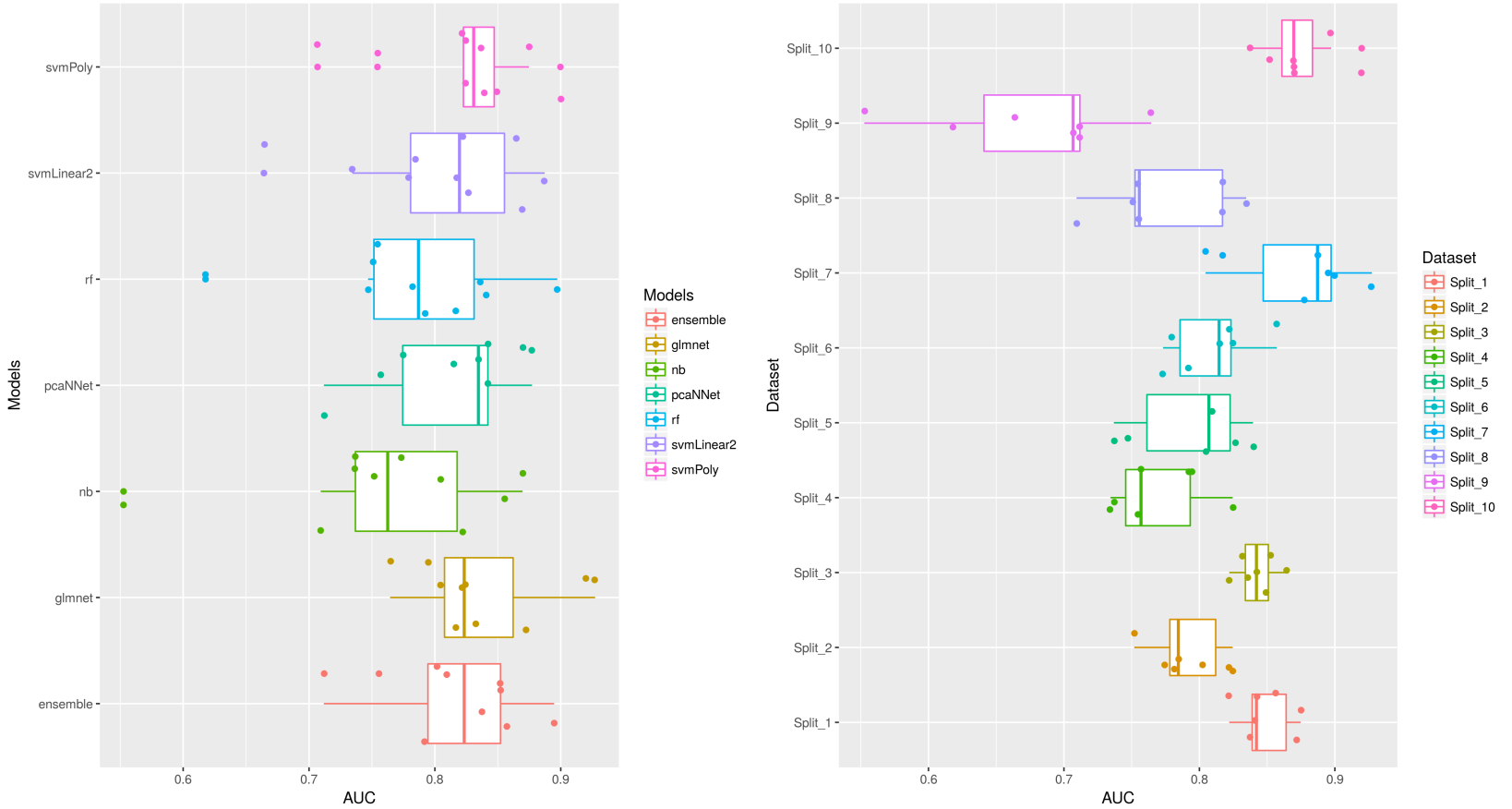
(b) RFE Single

Figure 5.2 Recursive Feature Elimination Results

Results of recursive feature elimination. (a) Shows individual linegraphs, and (b) shows linegraphs on the same scale for direct comparison. Y-axis shows predicted accuracy, X-axis shows number of features used.

The performance of the seven models across each of the 10 data splits was tested using corresponding test sets. This can be seen in Figure 5.3. The lowest AUC is observed in the 9th split (see Table 5.3) with an average AUC of 0.67 while the highest average AUC was found in split 10 with 0.88. The 5th data split failed to generate a pcaNNet.

Since the above analysis did not indicate that models in split 1 were outliers in any of the mentioned metrics, they were considered representative regarding overall performance. Models from split 1 were therefore used in further analysis, and features were examined. The top probe for the RF model was RBCK1, and the top probe for the GLMNET model was FUCA1. The GBM based ensemble model placed the highest weights on the svmPoly and RF models.



(a) AUC by Algorithm for all Models

(b) AUC by Data Split for all Models

Figure 5.3 AUC of Dejong models across 10 datasets, tested on internal validation

Figures exploring performance of 7 algorithms trained in 10 dejong datasets. All 70 algorithms were tested in internal test sets unseen by the algorithm in training. All 10 test sets consisted of 40 patient samples (21 Schizophrenia, 19 Controls). (a) Shows performance (AUC) of all 7 types of models. The 10 datapoints for each model type correspond to the 10 dejong training sets. (b) Shows the same data as in (a) but split across the 10 datasets, each with 7 data points representing a type of model.

Table 5.3 AUC for all 70 algorithms in Dejong test data

Dejong Test Data	Machine Learning Model						
	svmLinear2	svmPoly	glmnet	nb	pcaNNet	rf	ensemble
Split_1	0.82	0.87	0.87	0.86	0.84	0.84	0.84
Split_2	0.78	0.82	0.82	0.75	0.77	0.78	0.8
Split_3	0.86	0.85	0.83	0.82	0.84	0.84	0.85
Split_4	0.73	0.82	0.79	0.74	0.76	0.75	0.79
Split_5	0.83	0.84	0.8	0.74	NA	0.75	0.81
Split_6	0.78	0.82	0.82	0.77	0.81	0.79	0.86
Split_7	0.89	0.9	0.93	0.8	0.88	0.82	0.89
Split_8	0.82	0.75	0.82	0.71	0.83	0.75	0.76
Split_9	0.66	0.71	0.76	0.55	0.71	0.62	0.71
Split_10	0.87	0.84	0.92	0.87	0.87	0.9	0.85

Table of AUC for models tested on internal dejong test data. All rows correspond to AUC results of models trained on different initial splits of dejong data and tested on corresponding training data (n= 40).

Table 5.4 AUC for all 70 algorithms tested in GAP data

GAP Test Data	Machine Learning Model						
	svmLinear2	svmPoly	glmnet	nb	pcaNNet	rf	ensemble
Split_1	0.67	0.69	0.69	0.66	0.68	0.66	0.68
Split_2	0.67	0.67	0.68	0.66	0.69	0.65	0.67
Split_3	0.64	0.67	0.67	0.66	0.67	0.66	0.66
Split_4	0.68	0.67	0.67	0.65	0.67	0.64	0.67
Split_5	0.66	0.68	0.65	0.66	NA	0.66	0.66
Split_6	0.68	0.66	0.69	0.66	0.68	0.66	0.67
Split_7	0.64	0.67	0.67	0.67	0.67	0.66	0.65
Split_8	0.67	0.68	0.67	0.66	0.66	0.66	0.67
Split_9	0.67	0.66	0.7	0.67	0.68	0.66	0.68
Split_10	0.65	0.68	0.68	0.67	0.65	0.67	0.67

Table of AUC for models tested on GAP data. All rows correspond to AUC results of models trained on different initial splits of dejong data and tested on corresponding GAP data (n= 280).

5.3.3 Predicting FEP using Chronic Schizophrenia classifiers

GAP first-episode psychosis samples were predicted using all 70 models, with the AUC ranging from 0.65 to 0.7 across all splits and model types (see Table 5.4). Split 1 was used for analysis after this point for multiple reasons, since the results were robust across all splits.

Balanced Accuracy was calculated for all seven models in Split 1, for the full data, the schizophrenia samples and other psychosis samples (see Table 5.5). The highest balanced accuracy for schizophrenia was 67% (GLMNET model). The Ensemble model performed best for the full dataset and the other psychosis subset, with balanced accuracies of 64% and 61% respectively.

All models predicted class based on schizophrenia probability between 0 and 1. The cut-off for classification was 0.5, meaning all samples with a probability between 0.5 and 1 were assigned to the schizophrenia group, while all other samples were classed as controls. A perfect classifier would assign all controls a probability close to 0 and all FEP a probability close to 1. Density plots of all models showing the distribution probability for schizophrenia samples, other psychoses, controls and FEP can be found in (Appendix B: Figure B.1).

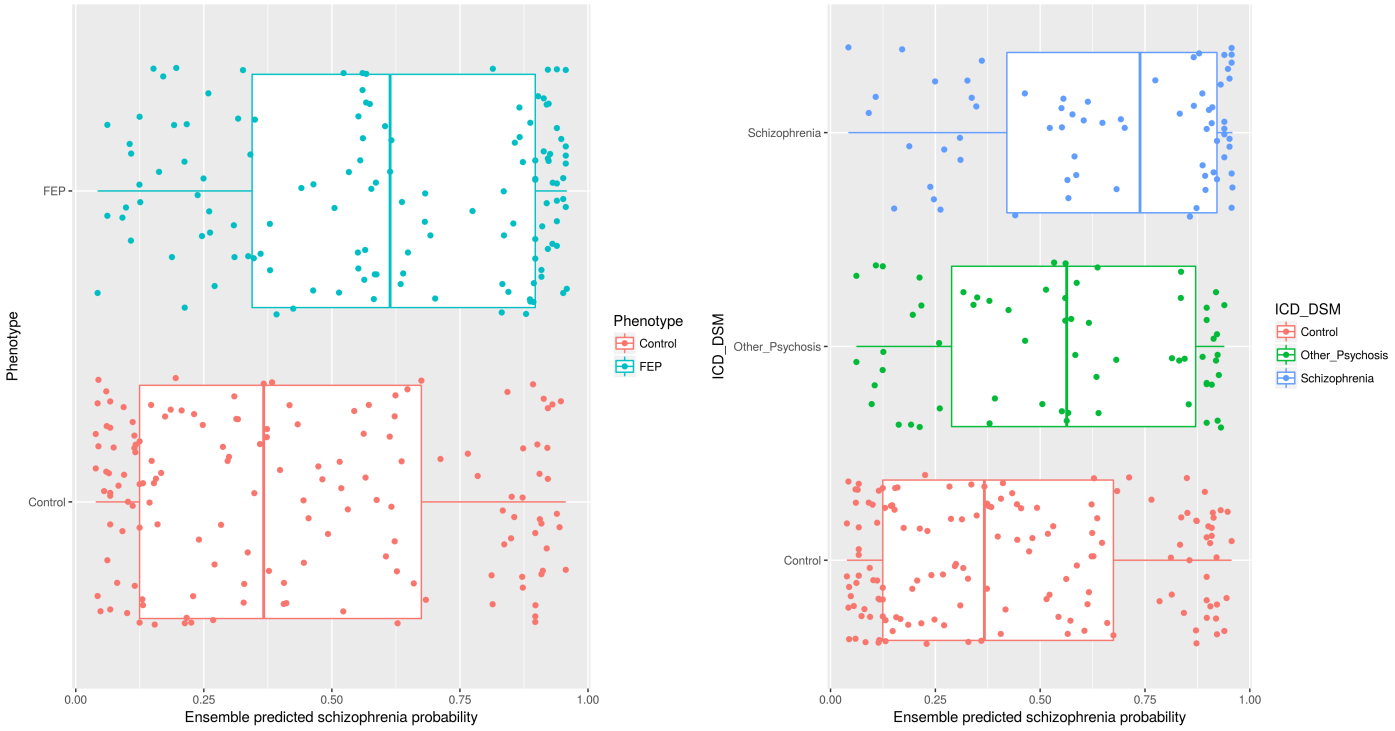
Boxplots of the final Ensemble model can be seen in Figure 5.4 where samples groups are compared based on diagnosis. For the Control-FEP comparison (Figure 5.4a) the median predicted probability (of belonging into the schizophrenia group) for FEP samples is 0.61 while for control samples it is 0.37. The difference between Control and FEP is statistically significant (p-value = 6.5e-07). For the Schizophrenia vs. Other Psychosis vs. Control comparison (Figure 5.4b) the median probability of schizophrenia assignment, as determined by the ensemble mode, is 0.74 for Schizophrenia samples, while it is 0.56 for other psychosis samples. ANOVA on the 3 sample groups identified a significant differences between the probabilities of Control and Schizophrenia samples (adj. p-value = 7e-07), Control and Other Psychosis (adj. p-value = 0.012), but not between Schizophrenia and Other Psychosis (adj. p-value = 0.12).

As such an ensemble model built on chronic schizophrenia, showed a significant difference in predictions for the control group compared to the FEP schizophrenia group, and the FEP other psychosis group. No significant difference was found between the Schizophrenia and other psychosis group.

Table 5.5 Analysis of Classification predictions in GAP data

GAP Data	Trained on	Method	Accuracy	Balanced Acc	Sensitivity	Specificity
Full	Split_1	svmLinear2	0.607	0.61	0.649	0.57
		svmPoly	0.625	0.626	0.641	0.611
		glmnet	0.632	0.635	0.679	0.591
		nb	0.611	0.611	0.611	0.611
		pcaNNet	0.611	0.614	0.664	0.564
		rf	0.611	0.61	0.595	0.624
		ensemble	0.639	0.641	0.672	0.611
Scz vs. Con	Split_1	svmLinear2	0.613	0.638	0.706	0.57
		svmPoly	0.641	0.658	0.706	0.611
		glmnet	0.641	0.67	0.75	0.591
		nb	0.622	0.629	0.647	0.611
		pcaNNet	0.613	0.642	0.721	0.564
		rf	0.631	0.636	0.647	0.624
		ensemble	0.645	0.666	0.721	0.611
Other vs. Con	Split_1	svmLinear2	0.575	0.579	0.587	0.57
		svmPoly	0.599	0.591	0.571	0.611
		glmnet	0.594	0.597	0.603	0.591
		nb	0.599	0.591	0.571	0.611
		pcaNNet	0.575	0.583	0.603	0.564
		rf	0.599	0.582	0.54	0.624
		ensemble	0.613	0.615	0.619	0.611

The table shows Accuracy, Sensitivity and Specificity for Predictions of 3 GAP subsets. Predictions were made by all Models created in by the first Dejong Split. The full GAP data (n= 280) is predicted based on the 2 category Dejong classifier. The second set includes only Psychosis samples that had an ICD or DSM diagnosis of Schizophrenia at some point, while the third GAP data subset excludes all samples that had a Schizophrenia Diagnosis at any point.



(a) Control vs. FEP

(b) Control vs. Scz vs. Other Psychosis

Figure 5.4 Boxplots of Ensemble Prediction Probabilities in GAP data

Boxplots of Ensemble probabilities in GAP data. The ensemble model was trained in the first dejong fold. Values are between 0 and 1, with higher values representing a higher confidence that the sample belongs to a patient. (a) Shows data split by Control and First Episode Psychosis (FEP). Differences between Control and FEP are statistically significant (p-value = 6.5e-07). (b) Shows the same data, split with first episode psychosis samples split across Schizophrenia samples, and Other Types of Psychosis. Differences are significant between Control and SCZ (adj. p-value = 7e-07), Control and Other Psychosis (adj. p-value = 0.012), but not between SCZ and Other Psychosis (adj. p-value = 0.12). The y-axis variation for individual points within a group, is unrelated to the data, and is purely a visual aid to clarify the distribution of points across the x-axis.

GAP misclassification

To look at misclassification in more detail, we looked at three main factors, Ethnicity, Gender and Medication. Ethnicity, Gender and Medication did not have a statistically significant effect on predictive confidence in the ensemble model. Boxplots showing the ensemble model stratified by the above factors can be seen in Appendix B: Figure B.2. Gender was close to being significant (p-value = 0.06), and this might be due to the Gender difference found in the training data (p-value < 0.001, see Table 5.2).

In addition we also tested PANSS subscales for significant correlation with predictive confidence in the ensemble model. This was done separately for Schizophrenia and Other Psychosis. None reached statistical significant, although the lowest p-value was for the positive subscale in schizophrenia samples (cor = 0.25, p-value = 0.07, n = 49).

Full stratification of DSM-IV and ICD-10 Diagnoses

Full stratification of DSM-IV and ICD-10 criteria was also performed for the FEP group (see Appendix B: Figure B.3). While the groups become to small for statistical analysis, they are noted for completeness. Thee categories had a median above 0.75 are Schizophrenia (median = 0.86, n = 32), Psychotic Disorder NOS (median = 0.77, n = 13), and Schizoaffective Disorder Depressed (median = 0.84, n = 9). The other categories include Mania with Psychosis (median = 0.56, n= 20), Schizophreniform Disorder (median = 0.56, n=27), followed by Schizoaffective Disorder Bipolar (median = 0.51, n = 7), and delusional disorder (median = 0.30, n= 7). The two patients with a diagnosis of Major Depression were both classed as Schizophrenic.

Out of the 68 patients classed as schizophrenic in either ICD10 or DSMIV, only 32 were diagnosed as Schizophrenic in the DSM-IV, while 25 were diagnosed with schizophreniform disorder, a diagnosis that is given to patients who do not meet the full criteria for schizophrenia according to the DSMIV. Interestingly, Schizophreniform samples show a wider range of predictions with a median predictive probability of 0.60 and a sensitivity of 0.64 (n = 25), while the 30 Schizophrenia cases have a median predictive probability of 0.86 and a sensitivity of 0.81 (n = 32). While a substantial amount of Schizophreniform diagnosed patients develop Schizophrenia eventually, their symptoms by definition have persisted for a shorter time period.

5.4 Discussion

5.4.1 Model Creation and Robustness

After model creation, we looked at all 70 models in internal test sets of DeJong. The ensemble models outperformed the strongest constitute model in just one data split. However the sample size for test samples in all cases was n= 40, and single outliers are bound to have a large impact at this point. Despite this, the average AUC for the Dejong training data was found to be 0.8. Which is in line with the results from other studies (Hess et al., 2016; Perkins et al., 2015; Takahashi et al., 2010), where an estimated AUC of 70%-86% was common.

5.4.2 Validation of Classifiers in GAP

The accuracy of chronic schizophrenia models, when applied to the GAP data was comparatively low. It was found that overall balanced accuracy was between 0.61 (svmLinear2 and rf) and 0.641 (ensemble). Even though the models were trained on Chronic Schizophrenia and not first episode psychosis, the accuracy is comparable and even outperforms (in the case of the ensemble model) the highest median estimates of bootstrapped GLMNET models (61% accuracy) constructed on GAP expression data (Chapter 4). It is important to keep in mind that GAP is both multi-ethnic and first episode cohort, while other studies have usually focused on chronic schizophrenia in a homogeneous population. One exception is the study by Hess et al. (2016), which included the Dejong dataset when training the machine learning algorithms. They achieved an AUC of 0.7 in testing data, which is comparable to the AUC of 0.64-0.7 achieved here.

5.4.3 Comparison of Schizophrenia and Other Psychoses

It may seem obvious that a classifier built on schizophrenia data would more accurately predict Schizophrenia in a second dataset, but the concept of Schizophrenia as a singular category is not as straightforward. Recent work shows considerable genetic overlap with conditions such as bipolar disorder and schizoaffective disorder (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013).

We found that predictive accuracy was greatest for Schizophrenia, followed by controls. While accuracy was lowest for Other Psychosis, they had a higher chance of being classed as Schizophrenic than controls. No significant difference in predictive confidence was found between Schizophrenia and Other Psychosis for the ensemble model. Also, but both groups revealed a statistically significant difference when compared with controls. This indicates,

in contrast with our hypothesis, and results from chapter 4, that a chronic schizophrenia signature classification model does show some predictive power in other psychoses.

These results show that it is possible to build a predictive model based on a white European chronic schizophrenia cohort and achieve significant predictive power in an ethnically mixed first episode psychosis population. While these models are not accurate enough to guide physicians in the clinic, they open the possibility of building more complex clinical models, by targeting subgroups, and integrating additional biomarkers.

5.4.4 Further diagnostic comparisons in First Episode Psychosis

Since schizophreniform is a less severe form of schizophrenia with better outcome, we compared the predictive power in patients with both diagnoses, according to DSM-IV records (see Appendix B: Figure B.3). The median predictive probability for Schizopreniform was of 0.60 (n = 25), while the Schizophrenia median predictive probability was 0.86 (n = 32). Since the classifier is based on chronic schizophrenia it makes sense that Schizophrenia patients would be more accurately predicted, and this might indicate that the classifier is using a signature in blood that is related to disease progression. However it might also be related to chronic medication use. Since this sample size is small, future analysis with more schizophreniform and schizophrenia patients would be necessary.

It is also interesting to note that the Schizoaffective disorder patients with depressed subtype were identified with higher confidence, than those with bipolar subtype. While the numbers are small for these groups, ICD-10 subgroups also show more accurate classification in line with more severe depressive symptoms. Delusions, Mania and the Schizoaffective Bipolar subtypes being predicted less accurately may be due to these disorders having more prominent elements of grandiosity, high self esteem and delusions of power.

It is possible that the some of this effect can be explained by stress that is common in depressive rumination, and schizophrenia and has a suppressive effect on the immune system in addition to numerous other long term consequences. Manic and Grandiose behaviour in delusional disorders if accompanied by a positive self perception may be objectivity damaging to the patient, but might not be perceived as such by them. This may also explain lower predictive power for patients with schizophreniform disorder as symptoms that have persisted for a shorter time period may not have had the same damaging effects and biological changes as would be caused by chronic stress. Including Duration of untreated Psychosis as a variable may be helpful in elucidating this further (Perkins et al., 2005).

5.4.5 Symptom Severity was not associated with Classification Accuracy

We did not find any significant relationship between PANSS and classification in either the Schizophrenia or other psychosis subsets. The most significant correlation was between positive symptoms and schizophrenia (cor = 0.25, p-value = 0.07, n = 49).

In our previous results we found that in a bootstrapped GLMNET model, the positive PANSS score correlated positively with the GAP Schizophrenia subset, but not with other other psychoses (Chapter 4). Recent research has identified a blood based gene expression signature (Jansen et al., 2016) in depression that shows immune suppression and activation in Major depression, and increased levels of IL-6. This mirrors the schizophrenia gene expression literature (Gardiner et al., 2013) and our own results (Chapter 3), were immune deregulation has been consistently identified.

While the results here are not significant, this may be related to low power due to small sample size. In addition positive symptoms in chronic schizophrenia are likely less severe (we did not have PANSS available for chronic patients at the time blood was taken), since they have been in treatment for a longer period. As such the blood signature related to positive symptoms, if it exists, would be reduced. From that perspective these results may be consistent with results from previous chapters, however this possibility has to be independently investigated.

5.4.6 Predictive probes are related to immunity and protein transport

GLMNET and Random Forest models provided a list of features used for predictions. The most significant of these for the GLMNET model was Fucosidase, Alpha-L- 1 (FUCA1) a protein implicated in lysosomal storage disease and glycoprotein metabolism. Lysosomal disruption has been observed previously in gene expression studies linking Schizophrenia and Bipolar disorder (Hess et al., 2016). This is of particular interest since lysosomal protein pathways play a role in signal transduction and regulation by virtue of up or downregulating receptors on the cell membrane.

The top probe for the random forest model was RBCK1, which is part of the LUBAC complex and plays a pivotal role in linear ubiquitination, all throughout the body. LUBAC is an essential protein in apoptosis and nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) signalling, by ubiquitinating upstream proteins facilitating their activation by phosporylation (Stieglitz et al., 2012; Tokunaga and Iwai, 2012). It plays an integral role in normal immune function.

Studies of schizophrenia and psychosis have found numerous links to genes apparently involved in immune signalling, cancer, metabolism and the cytoskeleten. While these pathways appear unrelated to psychiatric disorders, proteins in these pathways can have numerous roles, since endocytosis, exocytosis, apoptosis, cell migration and glucose metabolism are fundamental processes in normal neuronal functioning. These results further strengthen previous findings, albeit weakly.

5.4.7 Ensemble Model did not improve predictive power

Our ensemble model showed a modest improvement in classification , based on balanced accuracy, for first episode in general. Upon close examination this was due to a slight increase in sensitivity for non schizophrenia cases, from 0.603 to 0.619. Highest sensitivity for schizophrenia and overall balanced accuracy for schizophrenia cases was achieved by the glmnet model.

Further improvements may be achievable for the ensemble by increasing bootstrap iterations for the GBM based ensemble, using a different combinations of models and refining the feature selection process. In addition, GBM may not be ideal for a classification problem of this complexity. Boosting algorithms often perform poorly in datasets with a substantial amount of mislabeling, due to a high risk of overfitting. Since schizophrenia and psychosis are categories that primarily exist for historic and practical medical purposes, it is very likely that the best prediction approaches will need to take into account relatively high rates of misdiagnosis and mislabeling. One way of addressing this would be to use machine learning algorithms, that are more adept at handling mislabelled data.

Another option would be to train models on subsets that form more coherent sub categories, in this case data was trained on schizophrenia, but our data suggests that in first episode psychosis classification based on positive symptoms at least in schizophrenia may be useful.

To explore any of these questions with confidence, while avoiding the risk of over-fitting would require including additional data of comparable quality.

5.4.8 Clinical Application

Our results show similar performance to the recent large schizophrenia blood expression study by Hess et al. (2016). They report ROC of 0.9 - 0.99 in training data and ROC of 0.7-0.75 in test data. While our overall GAP dataset achieves an AUC of between 0.66 and 0.68, this is, as mentioned before, for the entire multi-ethnic First Episode Psychosis data. Performance increased for patients with a DSM-IV schizophrenia diagnosis, and also

patients with more depressed symptoms vs manic symptoms. We did not find any significant correlations between Schizophrenia samples and PANSS scores, in future studies it may be interesting to specifically contrast predictive power with patients who have bipolar disorder and specifically manic episodes.

5.4.9 Limitations

There is the possibility that our models simply identified unrelated markers that may correlate with some sort of stress rather than be directly associated with psychosis, something that has happened in the past. In order to address this, it would be important to test the model on gene expression data from other diseases, that are unrelated to psychosis. However even a signal that merely predicts general ill health, may in combination with other predictors such as PRS or family history, prove useful in a clinical setting.

5.4.10 Future work

Since psychosis is highly heterogeneous, and our results seem to indicate that Schizophrenia is more coherent than psychosis in general, finding clusters within psychosis seems like a more promising approach than treating first episode psychosis as a singular category. For this purpose Schizophrenia and perhaps schizo-affective disorder would form a category. Our results seem to indicate similar predictive accuracies for the schizo-affective subset as for schizophrenia. We also note a better prediction for patients with more severe diagnoses of depression. In both cases our sample size is limited and these results should be interpreted with extreme caution, but validation in gene expression data from depressed patients and patients with schizo-affective disorder, would be a natural progression. Previous findings have identified a gene expression signature for depression, that disappears in remission (Jansen et al., 2016).

This may simply indicate that patients with severe mental illness have a gene expression signature that is more related to metabolism and activity levels. This could explain why patients with manic symptoms are less likely to be classified correctly. In order to address this, drawing in additional clinical information about lifestyle and activity prior and during admission may be useful. In addition comparing results to those of patients with diseases, that produce similar changes in lifestyles, would help with establishing how specific the models are to mental health rather than general physical ill health.

As mentioned before further addition of genetic, metabolic and clinical data, should be incorporated in future work to investigate potential increases in performance. Unfortunately datasets of this complexity are difficult generate. However, several large scale project are

in the early stages of collecting data from thousands of participants, and will cover a range of biological omics data, activity data and clinical data long time periods. These include "the human project" in New York and "project baseline" at Duke University, which are both aiming to recruit 10,000 participants. When these projects mature they will provide the possibility of testing these models on a more general population.

Chapter 6

Discussion and Conclusions

6.1 Overview of the Thesis

In this thesis, I characterised the GAP gene expression data for the first time, to identify biomarkers and to work towards an empirical classification of first episode psychosis.

Chapter 1 introduced the complexity in diagnosing and treating psychosis sufferers appropriately, due to the multifaceted presentation, and the range of risk factors. I critically reviewed the literature on genetic and environmental factors and discussed findings in the context of a potential stress response model, which can account for gene-environment interactions. I introduced the concepts of transcriptomics and machine learning and examined the potential of these approaches in contributing to a more individualised medical approach.

Chapter 2 discusses methods and datasets used in this thesis. I introduced the GAP study and characterised the transcriptomic, clinical, demographic and genetic data used in this thesis. I described the methods employed in the preprocessing pipeline for the transcriptomic data, which represents the backbone of this thesis. Finally, I gave an overview of the approaches and algorithms relating to the machine learning aspects of this work and described how I implemented them.

Chapter 3 detailed the differential expression and network analysis experiments. The aim was to identify biomarkers and pathways to distinguish psychosis and control individuals. Enrichment of differentially expressed probes identified a significant association with pathways associated with transcription and viral infection. Further analysis using network approaches identified four modules related to first episode psychosis, all of which were highly enriched for brain-expressed genes and pathways. By including PANSS scores, I identified one additional module, which was associated with the severity of positive symptoms. This module was highly enriched for interferon gamma 1, viral and brain pathways.

Chapter 4 used bootstrapping and machine learning to generate a series of classification models and classification frequencies for samples, based on combinations of gene expression data, PRS and demographic variables. The aim was to estimate classification accuracy and characterise misclassification for future experiments. The classification accuracy was best for models using gene expression data but was limited to 0.61%. Examination of the data found that FEP samples with a later recorded schizophrenia diagnosis were more likely to be accurately classified than individuals who received other psychosis diagnoses. I also found a positive correlation between the positive sub-scale score of the PANSS and correct classification for people with a schizophrenia diagnosis, but again not for other psychoses. Transcripts chosen by the model were involved in immune function, iron regulation, mitochondrial function.

Chapter 5 details a machine learning based classification approach using an external gene expression dataset, first published by de Jong et al. (2012). The data was composed of samples taken from chronic schizophrenia patients and controls in the Netherlands. The aim was to build an ensemble model, based on multiple algorithms, to achieve higher classification accuracy and to examine the performance in the GAP FEP data. The accuracy of models was 80% in the Dutch cohort. The accuracy of FEP-control classification was 64% which was a small improvement over models trained on GAP. Sensitivity was found to be higher in GAP schizophrenia patients (0.65-0.75) than in other psychoses (0.59-0.62). The data also suggested that in other psychosis, depressive symptoms were classified more accurately than manic symptoms. Transcripts used in these models were again related to immune function.

In the remainder of this chapter, the implications and limitations of key findings of the work are discussed, before examining potential future directions.

6.2 Implications of key findings

6.2.1 Gene expression differences are consistent with a Stress response

Analysis of the GAP gene expression data, using differential expression and network approaches, to contrast FEP and controls, showed deregulation in pathways related to gene expression, viral infection, oxidative stress and innate immunity. These findings mirror the results from multiple other transcriptomic studies both in the case of schizophrenia (Gardiner et al., 2013; Hess et al., 2016) and Major depressive disorder (Jansen et al., 2016). These pathways and the direction of gene expression is consistent with a heightened stress response. The activation of the stress response can be modulated by a wide variety of biological, social,

and cultural factors, such as genetics, pathogen infections, physical injury, discrimination or neglect.

While the findings of this study are agnostic regarding the cause of the disruption of this pathway, it seems likely that all of the above-mentioned factors play a role to varying extents. Recent studies have implicated the importance of the HPA-axis in first episode psychosis, identifying associations between increased levels of cortisol and hippocampal volume (Mondelli et al., 2010b), as well as childhood trauma (Mondelli et al., 2010a). The findings presented here, are consistent with HPA-axis activation, and while this is not entirely novel, it strengthens existing literature, by providing further gene expression evidence in a large multi-ethnic first episode cohort.

6.2.2 Psychosis associated module is enriched for Schizophrenia risk genes

Following construction of modules using WGCNA, I performed enrichment analysis and included previously identified schizophrenia risk genes. I identified a large module highly enriched for several categories of schizophrenia risk genes. The categories used for enrichment were based on data compiled by Pirooznia et al. (2016), and included the genes from the psychiatric genetics consortium, rare copy-number variants, denovo mutations identified in sequencing studies, as well candidate genes and neuronal pathways determined to have a prior probability for schizophrenia risk.

While no evidence was found for an enrichment of the genes implicated by the recent PGC schizophrenia consortium, they were included in the composite list of all schizophrenia risk genes, which was the single most significant finding across all modules, and mapped to the Turquoise module. This module was also enriched for Postsynaptic density proteins, synaptome proteins, microglial markers, metal ion transport, mitochondrial function, Alzheimer's disease and glutamatergic synaptic function, as well as viral infections, platelet activation, vesicle transport and actin organisation.

While these pathways may seem fundamentally disconnected, they make sense in the context of HPA-axis activation. The release of cortisol suppresses the innate immune system and platelet activity, which use a lot of the same molecular machinery for endo and exocytosis that is required for neurotransmitter release. This relies in part on the actin cytoskeleton for vesicle transport. The stress response also suppresses the release of iron and other metals, to starve potential pathogens. Mitochondrial function is connected to apoptosis, nutrient starvation and oxidative stress. These perturbations are also found in Alzheimer's disease, where patients can also suffer from psychosis.

Glutamateterig synapses have been shown to play a major role in schizophrenia, highlighted by ketamine use and findings related to NMDAR autoimmunity, which both cause psychosis symptoms. Since this work is based on blood-based gene expression, these results are maybe better understood to affect pathways used in multiple systems, rather than indicating a glutamate system signature that can be detected in peripheral tissue. A high polygenic risk may sensitise both the glutamate system and pathways downstream of HPA-axis activation, by affecting the same genes used in both systems, which may explain differences in symptom presentation despite similar traumatic experiences. Alternatively, activation of the HPA axis may cause global changes in epigenetics that also affect brain expression.

It is important to note that this module was also correlated with BMI and Olanzapine (but not Risperidone). While no evidence was found for a significant contribution of antipsychotics, these results should be viewed with caution.

6.2.3 Positive Symptoms correlate with innate immune modules

I used available data from patients who had completed the Positive and Negative Syndrome Scale to identify WGCNA modules associated with symptom severity. The "greenyellow" module was the most significant finding, and was found to be positively correlated with positive symptoms, and negatively correlated with negative symptoms. The module was not correlated with the overall PANSS, psychosomatic symptoms, or case-control status.

After performing Gene enrichment analysis, I found significant enrichment for pathways related to type I interferon signalling, virus infections, as well as brain regions like the Hypothalamus. Expression differences correlated with severity of positive symptoms. Negative and positive symptoms are negatively correlated, and this may point towards a molecular mechanism that is different in the two dimensions.

One notable finding was that ADAR expression was negatively correlated with negative symptom severity in the greenyellow module. To account for a potential confounding effect of antipsychotic medication, I analysed medication free samples and found that ADAR was the most significantly correlated probe, in the greenyellow module, for positive symptoms, and was among the most highly correlated ones in negative symptoms.

ADAR is a ubiquitously expressed gene, with a variety of functions in innate immunity, cancer and neurological disorders (Mannion et al., 2015). It codes for an RNA editing protein with the primary function of marking and differentiating host RNA from viral RNA.

Deletion of ADAR is lethal, and downregulation causes severe autoimmunity as a ubiquitous antiviral response is induced via the Interferon I pathway, leading to upregulation of TNF-α and IL-6 among others. This upregulation has been consistently shown in gene expression studies for depression (Jansen et al., 2016) and schizophrenia (Gardiner et al.,

2013; Hess et al., 2016) and a large study looking at PTSD also recently found upregulation of the Interferon I pathway (Breen et al., 2017).

ADAR is located on chromosome 1 in the q21.1, which was identified by the International Schizophrenia Consortium as a rare deletion increasing Schizophrenia risk (International Schizophrenia Consortium, 2008).

As mentioned ADAR has also been implicated in various cancers (Mannion et al., 2015). Schizophrenia patients notably have lower rates of certain cancers (Ji et al., 2013) particularly prostate cancer, and ADAR suppression interestingly reduces prostate cancer proliferation in vitro and in vivo, (Salameh et al., 2015). Risk of cancer according to one study rises after successful treatment for patients (Ji et al., 2013). This may represent a selective advantage under certain circumstances, and partially explain differences in cancer rates.

The blood signature also shows a clearer correlation with positive symptoms than negative ones, which is supported by the machine learning results in chapter 3, that show a higher classification accuracy for schizophrenia patients as positive symptom severity increases. One interpretation of this might be that Positive symptoms are more likely to activate or be modulated by the HPA-axis and thus show a clearer signature in blood.

Overall these results show an interesting divergence between the Interferon I signature in positive and negative symptom severity, that may be worth exploring further. These findings are also consistent with a psychosis model that is influenced by environmental stressors.

6.2.4 Schizophrenia is more accurately predicted than other psychoses

The machine learning models I built, achieved comparably low accuracy between 61% and 64%. It was however notable that patients who received a schizophrenia diagnosis were more accurately identified compared to patients with other diagnoses. Interestingly I found that patients with affective disorders tend to show better classification accuracies if they had depressive symptoms rather than mania. In schizophrenia, high positive symptom severity increased classification accuracy as mentioned above. Evidence in the literature suggests that Schizophrenia and Bipolar disorder share a close genetic relationship, perhaps this is modulated more via depressive symptoms than manic ones.

This information may be useful in building future machine learning models for classification purposes and may help in identifying patient subgroups. Of note is that both machine learning approaches that relied on internal and external data, relied on probes that are deeply rooted in pathways associated with innate immunity, stress and the HPA-axis. This further strengthens the idea that HPA-axis activation is related to psychosis, at least in schizophrenia patients with positive symptoms.

While these results are not currently interesting from a clinical perspective, they may contribute in guiding a more targeted approach in the future. A simple binary classification approach was used, and this data suggests that using positive symptom scores rather than case-control status, might lead to better results.

6.3 Limitations

The GAP gene expression data is complex, including patients from different ethnic backgrounds, on different anti-psychotic medications and with a range of symptoms and diagnoses of varying degrees of severity.

As such the work presented in this thesis has multiple limitations. Since all chapters rely heavily on gene expression data in blood, there is a possibility of confounding with an endless number of variables. While attempts were made to account for unknown variables, the heterogeneous nature of the FEP sample could plausibly interfere in finding such signatures.

In addition anti-psychotic medication and smoking status likely need to be controlled for in a more comprehensive manner. Posthoc analyses, and correlations with modules were performed and compared with anti-psychotic free patients. However, only 17 of the FEP patients did not use antipsychotics, as such any lack of association should be viewed with appropriate scepticism.

Such subset analysis, was also performed when using PANSS scores and Diagnosis. This is subject to a lot of researchers degrees of freedom, and can introduce significant bias by testing many hypotheses and reporting only significant ones. To avoid this I aimed for transparency regarding analysis that was performed posthoc. No attempt was made to adjust for the number alternative hypotheses tested, but in many cases methods this would results in inadequate statistical hypothesis testing and would need to be independently verified in separate datasets regardless. This does highlight the requirement for cautious interpretation of these results. However analysis on subsets of potential confounding factors, that could not be included in the main analysis, are known a priori. As such they should be viewed less critically.

The weak findings in regard to PRS should be seen as preliminary since one third of samples came from individuals of African descent, where PRS has little to no predictive power. In addition imputation was used to predict PRS for 37 individuals, which further complicates interpretation. Combining this data with other datasets will be necessary to address this issue.

Another key limitation is that processing the data can introduce biases. Processing was performed from a perspective of Controls and First Episode Psychosis as distinct groups,

this was done to simplify analysis, but the results of the work presented here suggest that it might be important to conduct further analysis by separating groups by symptom severity and symptom type in the future. This is a limitation that is not unique to this work, and the fact that the data here suggests such a subdivision is useful in itself.

6.4 Future Directions

To address the issues discussed above, and test hypotheses generated, it would be important to include additional transcriptomic datasets. Hess et al. (2016) recently performed a large mega analysis of all publicly available expression data for schizophrenia, which largely mirrors the findings presented here. Since findings in Depression (Jansen et al., 2016) and PTSD mirror these results (Breen et al., 2017), incorporating all publicly available gene expression datasets from a range of psychiatric disorders in a comprehensive analysis would be natural progression to elucidate commonalities and differences. Unfortunately few datasets have the breadth of available data that is found in the GAP study.

As such further work internal to the GAP study is extensive. Additional biological data that includes mass spectrometry based proteomics, methylation and neuroimaging, as well to 3 and 12 month followup data for transcriptomics and neuroimaging in a subset of participants can be incorporated in the future. This would allow the testing of a series of hypothesis discussed in this thesis. It would also allow for more complex classification models. Finally it would allow for the generation of expression and protein quantitative trait loci (eQTL and pQTL) which could also be incorporated into potential classifiers.

An important step would be to incorporate data on trauma, drug use and other risk factors, that may help with generating better predictive models and exploring a potential link with the HPA-axis.

6.5 Concluding Remarks

The results of this thesis add to the existing literature which strongly implicates innate immunity and stress in its broadest definition as a contributing factor to the development of psychosis. While this is likely a very non-specific indicator of psychosis risk at the molecular level, further studies may identify a more specific "sub-signature" in certain patient groups. The results presented here showed some indication that positive symptom severity was linked to the strength of interferon I pathway deregulation for example, and predictive modelling approaches indicated a more coherent blood based expression signature for schizophrenia suffers.

Taking this further, it may be possible to ultimately identify a gene expression signature in blood that specifically points towards high HPA-axis activation via rumination or anxiety, or identify changes that are more closely linked to a stress response mediated by the NF-κB pathway, due to environmental exposures or pathogens. Genetics, Epigenetic, patient history and living conditions would likely help elucidate this further, allowing for a more targeted and personalised intervention as is seen in the case of Anti-NMDAR encephalitis.

The renowned author and neuroscientist Robert Sapolsky, in his best-selling book "Why Zebras don't get ulcers", once explained that a Zebra leads a very relaxing social live in a herd, punctuated only by short bursts of stress when fleeing from a predator, before succumbing to being eaten or returning to a presumably satisfying existence of eating grass. The human stress response evolved in same context, not predicting that we would leave our distant cousins behind, to build cities and societies; filled with inequality, discrimination and sleepless nights in front of excel spreadsheets, which one cannot outrun.

In this context, an easily stimulated stress response, whether it is caused by genetic or environmental factors, can quickly become maladaptive. Future study of how the stress response behaves in modern society, for individuals of varying genetic backgrounds, can perhaps help in shaping social policies and interventions that allow people to thrive despite our long outdated biology.

References

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS ONE*, 4(7):e6098.

Aboraya, A., Tien, A., Stevenson, J., and Crosby, K. (1989). Schedules for Clinical Assessment in Neuropsychiatry (SCAN): introduction to WV's mental health community. *The West Virginia medical journal*, 94(6):326–8.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. volume 1973, pages 420–434. Springer Verlag.

Allen, J. D., Chen, M., and Xie, Y. (2009). Model-Based Background Correction (MBCB): R Methods and GUI for Illumina Bead-array Data. *Journal of cancer science & therapy*, 1(1):25–27.

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175–185.

American Psychiatric Association. and American Psychiatric Association. DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5*. American Psychiatric Association, Arlington VA ;Washington D.C., 5th ed. edition.

Anderson, G. and Maes, M. (2012). Melatonin: An overlooked factor in schizophrenia and in the inhibition of anti-psychotic side effects.

Aranow, C. (2011). Vitamin D and the immune system. *Journal of investigative medicine : the official publication of the American Federation for Clinical Research*, 59(6):881–6.

Arion, D., Unger, T., Lewis, D. A., Levitt, P., and Mirnics, K. (2007). Molecular Evidence for Increased Expression of Genes Related to Immune and Chaperone Function in the Prefrontal Cortex in Schizophrenia. *Biological Psychiatry*, 62(7):711–721.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9.

Awad, A. G. and Voruganti, L. N. P. (2008). The burden of schizophrenia on caregivers: A review.

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Research*, 33(18):5914–5923.

Bebbington, P. and Nayani, T. (1996). The psychosis screening questionnaire. *International Journal of Methods in Psychiatric Research*, 5(1):11–19.

Breen, M. S., Tylee, D. S., Maihofer, A. X., Neylan, T. C., Mehta, D., Binder, E., Chandler, S. D., Hess, J. L., Kremen, W. S., Risbrough, V. B., Woelk, C. H., Baker, D. G., Nievergelt, C. M., Tsuang, M. T., Buxbaum, J. D., and Glatt, S. J. (2017). PTSD Blood Transcriptome Mega-Analysis: Shared Inflammatory Pathways Across Biological Sex and Modes of Trauma. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*.

Breiman, L. (2001). Random forests. 45(1):5–32.

Brown, H. E. and Roffman, J. L. (2014). Vitamin supplementation in the treatment of schizophrenia. *CNS drugs*, 28(7):611–22.

Cardno, A. G., Marshall, E. J., Coid, B., Macdonald, A. M., Ribchester, T. R., Davies, N. J., Venturi, P., Jones, L. a., Lewis, S. W., Sham, P. C., Gottesman, I. I., Farmer, A. E., McGuffin, P., Reveley, A. M., and Murray, R. M. (1999). Heritability Estimates for Psychotic Disorders. *Archives of General Psychiatry*, 56(2):162.

Caspi, A., Moffitt, T. E., Cannon, M., McClay, J., Murray, R., Harrington, H., Taylor, A., Arseneault, L., Williams, B., Braithwaite, A., Poulton, R., and Craig, I. W. (2005). Moderation of the Effect of Adolescent-Onset Cannabis Use on Adult Psychosis by a Functional Polymorphism in the Catechol-O-Methyltransferase Gene: Longitudinal Evidence of a Gene X Environment Interaction. *Biological Psychiatry*, 57(10):1117–1127.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. 20(3):273–297.

Craddock, R. M., Huang, J. T., Jackson, E., Harris, N., Torrey, E. F., Herberth, M., and Bahn, S. (2008). Increased alpha-defensins as a blood marker for schizophrenia susceptibility. *Molecular & cellular proteomics : MCP*, 7(7):1204–13.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, 381(9875):1371–1379.

de Jong, S., Boks, M. P. M., Fuller, T. F., Strengman, E., Janson, E., de Kovel, C. G. F., Ori, A. P. S., Vi, N., Mulder, F., Blom, J. D., Glenthøj, B., Schubart, C. D., Cahn, W., Kahn, R. S., Horvath, S., and Ophoff, R. a. (2012). A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PloS one*, 7(6):e39498.

Di Forti, M., Iyegbe, C., Sallis, H., Kolliakou, A., Falcone, M. A., Paparelli, A., Sirianni, M., La Cascia, C., Stilo, S. A., Marques, T. R., Handley, R., Mondelli, V., Dazzan, P., Pariante, C., David, A. S., Morgan, C., Powell, J., and Murray, R. M. (2012). Confirmation that the AKT1 (rs2494732) Genotype Influences the Risk of Psychosis in Cannabis Users. *Biological Psychiatry*, 72(10):811–816.

Di Forti, M., Marconi, A., Carra, E., Fraietta, S., Trotta, A., Bonomo, M., Bianconi, F., Gardner-Sood, P., O'Connor, J., Russo, M., Stilo, S. A., Marques, T. R., Mondelli, V., Dazzan, P., Pariante, C., David, A. S., Gaughran, F., Atakan, Z., Iyegbe, C., Powell, J., Morgan, C., Lynskey, M., and Murray, R. M. (2015). Proportion of patients in south London with first-episode psychosis attributable to use of high potency cannabis: a case-control study. *The lancet. Psychiatry*, 2(3):233–8.

Di Forti, M., Morgan, C., Dazzan, P., Pariante, C., Mondelli, V., Marques, T. R., Handley, R., Luzi, S., Russo, M., Paparelli, A., Butt, A., Stilo, S. A., Wiffen, B., Powell, J., and Murray, R. M. (2009). High-potency cannabis and the risk of psychosis. *The British journal of psychiatry : the journal of mental science*, 195(6):488–91.

Ding, L.-H., Xie, Y., Park, S., Xiao, G., and Story, M. D. (2008). Enhanced identification and biological validation of differential gene expression via Illumina whole-genome expression arrays through the use of the model-based background correction methodology. *Nucleic Acids Research*, 36(10).

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24:1547–8.

Dunning, M., Lynch, A., and Eldridge, M. (2015). illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4).

Efron, B. (2013). Bayes' theorem in the 21st century.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. 1(1):54–75.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15.

Fosse, R., Joseph, J., and Jones, M. (2016). Schizophrenia: A critical view on genetic effects. 8(1):72–84.

Fosse, R., Joseph, J., and Richardson, K. (2015). A critical assessment of the equal-environment assumption of the twin method for schizophrenia. *Frontiers in psychiatry*, 6:62.

Frank, S., Peters, M. A., Wehmeyer, C., Strietholt, S., Koers-Wunrau, C., Bertrand, J., Heitzmann, M., Hillmann, A., Sherwood, J., Seyfert, C., Gay, S., and Pap, T. (2013). Regulation of matrixmetalloproteinase-3 and matrixmetalloproteinase-13 by SUMO-2/3 through the transcription factor NF-κB. *Ann Rheum Dis*, 72:1874–1881.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Ganz, T. (2009). Iron in innate immunity: starve the invaders.

Gardiner, E. J., Cairns, M. J., Liu, B., Beveridge, N. J., Carr, V., Kelly, B., Scott, R. J., and Tooney, P. a. (2013). Gene expression analysis reveals schizophrenia-associated dysregulation of immune pathways in peripheral blood mononuclear cells. *Journal of psychiatric research*, 47(4):425–37.

Gaujoux, R. and Seoighe, C. (2013). CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics (Oxford, England)*, 29(17):2211–2.

Goetz, D. H., Holmes, M. A., Borregaard, N., Bluhm, M. E., Raymond, K. N., and Strong, R. K. (2002). The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition. 10(5):1033–1043.

Guest, P. C., Chan, M. K., Gottschalk, M. G., and Bahn, S. (2014). The use of proteomic biomarkers for improved diagnosis and stratification of schizophrenia patients. *Biomarkers in medicine*, 8(1):15–27.

Gusev, A., Lee, S., Trynka, G., Finucane, H., Vilhjálmsson, B., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., Kähler, A. K., Hultman, C. M., Purcell, S. M., McCarroll, S. A., Daly, M., Pasaniuc, B., Sullivan, P. F., Neale, B. M., Wray, N. R., Raychaudhuri, S., Price, A. L., Ripke, S., Neale, B., Corvin, A., Walters, J., Farh, K.-H., Holmans, P., Lee, P., Bulik-Sullivan, B., Collier, D., Huang, H., Pers, T., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S., Begemann, M., Belliveau, R., Bene, J., Bergen, S., Bevilacqua, E., Bigdeli, T., Black, D., Børglum, A., Bruggeman, R., Buccola, N., Buckner, R., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R., Carr, V., Carrera, N., Catts, S., Chambert, K., Chan, R., Chen, R., Chen, E., Cheng, W., Cheung, E., Chong, S., Cloninger, C., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J., Curtis, D., Davidson, M., Davis, K., Degenhardt, F., Del Favero, J., DeLisi, L., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A., Farrell, M., Frank, J., Franke, L., Freedman, R., Freimer, N., Friedl, M., Friedman, J., Fromer, M., Genovese, G., Georgieva, L., Gershon, E., Giegling, I., Giusti-Rodrguez, P., Godard, S., Goldstein, J., Golimbet, V., Gopal, S., Gratten, J., Grove, J., de Haan, L., Hammer, C., Hamshere, M., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A., Henskens, F., Herms, S., Hirschhorn, J., Hoffmann, P., Hofman, A., Hollegaard, M., Hougaard, D., Ikeda, M., Joa, I., Julià, A., Kahn, R., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M., Kelly, B., Kennedy, J., Khrunin, A., Kim, Y., Klovins, J., Knowles, J., Konte, B., Kucinskas, V., Kucinskiene, Z., Kuzelova-Ptackova, H., Kähler, A., Laurent, C., Keong, J., Lee, S., Legge, S., Lerer, B., Li, M., Li, T., Liang, K.-Y., Lieberman, J., Limborska, S., Loughland, C., Lubinski, J., Lnnqvist, J., Macek, M., Magnusson, P., Maher, B., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R., McDonald, C., McIntosh, A., Meier, S., Meijer, C., Melegh, B., Melle, I., Mesholam-Gately, R., Metspalu, A., Michie, P., Milani, L., Milanova, V., Mokrab, Y., Morris, D., Mors, O., Mortensen, P., Murphy, K., Murray, R., Myin-Germeys, I., Mller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D., Nestadt, G., Nicodemus, K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F., Oh, S.-Y., Olincy, A., Olsen, L., Van Os, J., Pantelis, C., Papadimitriou, G., Papiol, S., Parkhomenko, E., Pato, M., Paunio, T., Pejovic-Milovancevic, M., Perkins, D., Pietilinen, O., Pimm, J., Pocklington, A., Powell, J., Price, A., Pulver, A., Purcell, S., Quested, D., Rasmussen, H., Reichenberg, A., Reimers, M., Richards, A., Roffman, J., Roussos, P., Ruderfer, D., Salomaa, V., Sanders, A., Schall, U., Schubert, C., Schulze, T., Schwab, S., Scolnick, E., Scott, R., Seidman, L., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J., Sim, K., Slominsky, P., Smoller, J., So, H.-C., Spencer, C., Stahl, E., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R., Strengman, E., Strohmaier, J., Stroup, T., Subramaniam, M., Suvisaari, J., Svrakic, D., Szatkiewicz, J., Sderman, E., Thirumalai, S., Toncheva, D.,

Tooney, P., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B., Weiser, M., Wildenauer, D., Williams, N., Williams, S., Witt, S., Wolen, A., Wong, E., Wormley, B., Wu, J., Xi, H., Zai, C., Zheng, X., Zimprich, F., Wray, N., Stefansson, K., Visscher, P., Adolfsson, R., Andreassen, O., Blackwood, D., Bramon, E., Buxbaum, J., Brglum, A., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P., Gill, M., Gurling, H., Hultman, C., Iwata, N., Jablensky, A., Jönsson, E., Kendler, K., Kirov, G., Knight, J., Lencz, T., Levinson, D., Li, Q., Liu, J., Malhotra, A., McCarroll, S., McQuillin, A., Moran, J., Mortensen, P., Mowry, B., Nthen, M., Ophoff, R., Owen, M., Palotie, A., Pato, C., Petryshen, T., Posthuma, D., Rietschel, M., Riley, B., Rujescu, D., Sham, P., Sklar, P., St. Clair, D., Weinberger, D., Wendland, J., Werge, T., Daly, M., Sullivan, P., O'Donovan, M., Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., Bergen, S., Magnusson, P. K., Neale, B. M., Ruderfer, D., Scolnick, E., Purcell, S., McCarroll, S., Sklar, P., Hultman, C. M., and Sullivan, P. F. (2014). Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *The American Journal of Human Genetics*, 95(5):535–552.

Häfner, H. (2003). Gender differences in schizophrenia. *Psychoneuroendocrinology*, 28(SUPPL. 2):17–54.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Hess, J. L., Tylee, D. S., Barve, R., de Jong, S., Ophoff, R. A., Kumarasinghe, N., Tooney, P., Schall, U., Gardiner, E., Beveridge, N. J., Scott, R. J., Yasawardene, S., Perera, A., Mendis, J., Carr, V., Kelly, B., Cairns, M., Tsuang, M. T., and Glatt, S. J. (2016). Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia. *Schizophrenia Research*, 176(2-3):114–124.

Horváth, S. and Mirnics, K. (2015). Schizophrenia as a Disorder of Molecular Pathways. *Biological Psychiatry*, 77(1):22–28.

Howes, O. D. and Murray, R. M. (2014). Schizophrenia: an integrated sociodevelopmental-cognitive model. *Lancet (London, England)*, 383(9929):1677–1687.

Iavarone, F., Melis, M., Platania, G., Cabras, T., Manconi, B., Petruzzelli, R., Cordaro, M., Siracusano, A., Faa, G., Messana, I., Zanasi, M., and Castagnola, M. (2014). Characterization of salivary proteins of schizophrenic and bipolar disorder patients by top-down proteomics. *Journal of proteomics*, 103:15–22.

International Schizophrenia Consortium, T. I. S. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–41.

Iyegbe, C., Campbell, D., Butler, A., Ajnakina, O., and Sham, P. (2014). The emerging molecular architecture of schizophrenia, polygenic risk scores and the clinical implications for GxE research. *Social psychiatry and psychiatric epidemiology*, 49(2):169–82.

Jakobsen, K. D., Frederiksen, J. N., Parnas, J., Werge, T., and Jakobsen, K. D. (2006). Diagnostic Agreement of Schizophrenia Spectrum Disorders among Chronic Patients with Functional Psychoses. *Psychopathology*, 3939.

Jansen, R., Penninx, B. W. J. H., Madar, V., Xia, K., Milaneschi, Y., Hottenga, J. J., Hammerschlag, A. R., Beekman, A., van der Wee, N., Smit, J. H., Brooks, A. I., Tischfield, J., Posthuma, D., Schoevers, R., van Grootheest, G., Willemsen, G., de Geus, E. J., Boomsma, D. I., Wright, F. A., Zou, F., Sun, W., and Sullivan, P. F. (2016). Gene expression in major depressive disorder. *Molecular Psychiatry*, 21(3):339–347.

Jb, K., Errazuriz A, Tj, C., C, M., Jackson D, Mccrone P, Rm, M., and Pb, J. (2012). Systematic Review of the Incidence and Prevalence of Schizophrenia and Other Psychoses in England EXECUTIVE SUMMARY Background (Chapter 1).

Jelenkovic, A., Yokoyama, Y., Sund, R., Honda, C., Bogl, L. H., Aaltonen, S., Ji, F., Ning, F., Pang, Z., Ordoñana, J. R., Sánchez-Romera, J. F., Colodro-Conde, L., Burt, S. A., Klump, K. L., Medland, S. E., Montgomery, G. W., Kandler, C., McAdams, T. A., Eley, T. C., Gregory, A. M., Saudino, K. J., Dubois, L., Boivin, M., Tarnoki, A. D., Tarnoki, D. L., Haworth, C. M. A., Plomin, R., Öncel, S. Y., Aliev, F., Stazi, M. A., Fagnani, C., D'Ippolito, C., Craig, J. M., Saffery, R., Siribaddana, S. H., Hotopf, M., Sumathipala, A., Rijsdijk, F., Spector, T., Mangino, M., Lachance, G., Gatz, M., Butler, D. A., Bayasgalan, G., Narandalai, D., Freitas, D. L., Maia, J. A., Harden, K. P., Tucker-Drob, E. M., Kim, B., Chong, Y., Hong, C., Shin, H. J., Christensen, K., Skytthe, A., Kyvik, K. O., Derom, C. A., Vlietinck, R. F., Loos, R. J. F., Cozen, W., Hwang, A. E., Mack, T. M., He, M., Ding, X., Chang, B., Silberg, J. L., Eaves, L. J., Maes, H. H., Cutler, T. L., Hopper, J. L., Aujard, K., Magnusson, P. K. E., Pedersen, N. L., Aslan, A. K. D., Song, Y.-M., Yang, S., Lee, K., Baker, L. A., Tuvblad, C., Bjerregaard-Andersen, M., Beck-Nielsen, H., Sodemann, M., Heikkilä, K., Tan, Q., Zhang, D., Swan, G. E., Krasnow, R., Jang, K. L., Knafo-Noam, A., Mankuta, D., Abramson, L., Lichtenstein, P., Krueger, R. F., McGue, M., Pahlen, S., Tynelius, P., Duncan, G. E., Buchwald, D., Corley, R. P., Huibregtse, B. M., Nelson, T. L., Whitfield, K. E., Franz, C. E., Kremen, W. S., Lyons, M. J., Ooki, S., Brandt, I., Nilsen, T. S., Inui, F., Watanabe, M., Bartels, M., van Beijsterveldt, T. C. E. M., Wardle, J., Llewellyn, C. H., Fisher, A., Rebato, E., Martin, N. G., Iwatani, Y., Hayakawa, K., Sung, J., Harris, J. R., Willemsen, G., Busjahn, A., Goldberg, J. H., Rasmussen, F., Hur, Y.-M., Boomsma, D. I., Sørensen, T. I. A., Kaprio, J., and Silventoinen, K. (2015). Zygosity Differences in Height and Body Mass Index of Twins From Infancy to Old Age: A Study of the CODATwins Project. *Twin Research and Human Genetics*, 18(05):557–570.

Ji, J., Sundquist, K., Ning, Y., Kendler, K. S., Sundquist, J., and Chen, X. (2013). Incidence of cancer in patients with schizophrenia and their first-degree relatives: a population-based study in Sweden. *Schizophrenia bulletin*, 39(3):527–36.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic acids research*, 32(Database issue):277–80.

Karanikas, E., Antoniadis, D., and Garyfallos, G. D. (2014). The Role of Cortisol in First Episode of Psychosis: A Systematic Review. *Current Psychiatry Reports*, 16(11):503.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kauffmann, A., Rayner, T. F., Parkinson, H., Kapushesky, M., Lukk, M., Brazma, A., and Huber, W. (2009). Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics (Oxford, England)*, 25(16):2092–4.

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin*, 13(2):261–76.

Kendler, S. (1983). AMERICAN on. (November):1413–1425.

Kirkbride, J., Coid, J. W., Morgan, C., Fearon, P., Dazzan, P., Yang, M., Lloyd, T., Harrison, G. L., and Murray, R. M. (2010). Europe PMC Funders Group Translating the epidemiology of psychosis into public mental health : evidence , challenges and future prospects. 9(2):4–14.

Kirsch, I. (2014). Antidepressants and the Placebo Effect. *Zeitschrift fur Psychologie*, 222(3):128–134.

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., and Johnson, B. T. (2008). Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration. *PLoS Medicine*, 5(2):e45.

Kirschbaum, C., Wüst, S., and Hellhammer, D. (1992). Consistent sex differences in cortisol responses to psychological stress. *Psychosomatic medicine*, 54(6):648–57.

Kobrynski, L. J. and Sullivan, K. E. (2007). Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes. *The Lancet*, 370(9596):1443–1452.

Korkmaz, S., Yıldız, S., Korucu, T., Gundogan, B., Sunbul, Z. E., Korkmaz, H., and Atmaca, M. (2015). Frequency of anemia in chronic psychiatry patients. *Neuropsychiatric disease and treatment*, 11:2737–41.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5).

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*.

Kumarasinghe, N., Tooney, P. a., and Schall, U. (2012). Finding the needle in the haystack: a review of microarray gene expression research into schizophrenia. *The Australian and New Zealand journal of psychiatry*, 46(7):598–610.

Kuzman, M. R., Medved, V., Terzic, J., and Krainc, D. (2009). Genome-wide expression analysis of peripheral blood identifies candidate biomarkers for schizophrenia. *Journal of psychiatric research*, 43(13):1073–7.

Laing, R. D., Esterson, A., and Mantel, H. (2016). *Sanity, madness and the family*. Taylor and Francis.

Lander, E. and Schork, N. (1994). Genetic dissection of complex traits. *Science*, 265(5181).

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9:559.

Lee, J., Goh, L.-K., Chen, G., Verma, S., Tan, C.-H., and Lee, T.-S. (2012). Analysis of blood-based gene expression signature in first-episode psychosis. *Psychiatry research*, 200(1):52–4.

Leek, J. and Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9).

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, 28(6):882–3.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, 11(10):733–9.

Lelli-Chiesa, G., Kempton, M. J., Jogia, J., Tatarelli, R., Girardi, P., Powell, J., Collier, D. A., and Frangou, S. (2011). The impact of the Val158Met catechol- O-methyltransferase genotype on neural correlates of sad facial affect processing in patients with bipolar disorder and their relatives. *Psychological Medicine*, 41(04):779–788.

Levin, Y., Wang, L., Schwarz, E., Koethe, D., Leweke, F. M., and Bahn, S. (2010). Global proteomic profiling reveals altered proteomic signature in schizophrenia serum. *Molecular Psychiatry*, 15(11):1088–1100.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., O'Donovan, M. C., Neale, B. M., Patterson, N., Price, A. L., and Price, A. L. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics*, 47(12):1385–1392.

Lu, Y.-F., Goldstein, D. B., Angrist, M., and Cavalleri, G. (2014). Personalized medicine and human genetic diversity. *Cold Spring Harbor perspectives in medicine*, 4(9):a008581.

Mah, L., Szabuniewicz, C., and Fiocco, A. J. (2016). Can anxiety damage the brain? *Current Opinion in Psychiatry*, 29(1):56–63.

Maher, B. (2008). The case of the missing heritability. *Nature*, 456(November):18–21.

Mannion, N., Arieti, F., Gallo, A., Keegan, L., and O'Connell, M. (2015). New Insights into the Biological Role of Mammalian ADARs; the RNA Editing Proteins. *Biomolecules*, 5(4):2338–2362.

Mayer, Z. (2017). caretEnsemble (developer version).

McCarty, D. E., Chesson, A. L., Jain, S. K., and Marino, A. A. (2014). The link between vitamin D metabolism and sleep medicine.

McGrath, J., Saha, S., Chant, D., and Welham, J. (2008). Schizophrenia: A concise overview of incidence, prevalence, and mortality.

Meyer, D. (2001). Support Vector Machines.

Middleton, F. A., Pato, C. N., Gentile, K. L., McGann, L., Brown, A. M., Trauzzi, M., Diab, H., Morley, C. P., Medeiros, H., Macedo, A., Azevedo, M. H., and Pato, M. T. (2005). Gene expression analysis of peripheral blood leukocytes from discordant sib-pairs with schizophrenia and bipolar disorder reveals points of convergence between genetic and functional genomic approaches. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 136B(1):12–25.

Miller, A. H. and Raison, C. L. (2016). The role of inflammation in depression: From evolutionary imperative to modern treatment target.

Minsky, M. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA.

Mistry, M., Gillis, J., and Pavlidis, P. (2013). Genome-wide expression profiling of schizophrenia using a large combined cohort. *Molecular psychiatry*, 18(2):215–25.

Mondelli, V., Dazzan, P., Hepgul, N., Di Forti, M., Aas, M., D'Albenzio, A., Di Nicola, M., Fisher, H., Handley, R., Marques, T. R., Morgan, C., Navari, S., Taylor, H., Papadopoulos, A., Aitchison, K. J., Murray, R. M., and Pariante, C. M. (2010a). Abnormal cortisol levels during the day and cortisol awakening response in first-episode psychosis: The role of stress and of antipsychotic treatment. *Schizophrenia Research*, 116(2-3):234–242.

Mondelli, V., Pariante, C. M., Navari, S., Aas, M., D'Albenzio, A., Di Forti, M., Handley, R., Hepgul, N., Marques, T. R., Taylor, H., Papadopoulos, A. S., Aitchison, K. J., Murray, R. M., and Dazzan, P. (2010b). Higher cortisol levels are associated with smaller left hippocampal volume in first-episode psychosis. *Schizophrenia research*, 119(1-3):75–8.

Moylan, S., Berk, M., Dean, O. M., Samuni, Y., Williams, L. J., O'Neil, A., Hayley, A. C., Pasco, J. A., Anderson, G., Jacka, F. N., and Maes, M. (2014). Oxidative & nitrosative stress in depression: Why so much stress?

Narayan, C. L., Shikha, D., and Shekhar, S. (2015). Schizophrenia in identical twins. *Indian journal of psychiatry*, 57(3):323–4.

Newhouse, S. J. (2013). Illumina expression workflow.

Oldham, M. C., Langfelder, P., and Horvath, S. (2012). Network methods for describing sample relationships in genomic datasets : application to Huntington's disease. *BMC Systems Biology*, 6(63).

O'Malley, A. J., Frank, R. G., and Normand, S.-L. T. (2011). Estimating cost-offsets of new medications: use of new antipsychotics and mental health costs for schizophrenia. *Statistics in medicine*, 30(16):1971–88.

Organization, W. H. (1992). The ICD–10 Classification of Mental and Behavioural Disorders. *Clinical Description and Diagnostic Guidelines*.

Paris, J. J., Franco, C., Sodano, R., Freidenberg, B., Gordis, E., Anderson, D. A., Forsyth, J. P., Wulfert, E., and Frye, C. A. (2010). Sex differences in salivary cortisol in response to acute stressors among healthy participants, in recreational or pathological gamblers, and in those with posttraumatic stress disorder. *Hormones and Behavior*, 57(1):35–45.

Parrow, N. L., Fleming, R. E., and Minnick, M. F. (2013). Sequestration and scavenging of iron in infection.

Penckofer, S., Kouba, J., Byrn, M., and Estwing Ferrans, C. (2010). Vitamin D and depression: where is all the sunshine? *Issues in mental health nursing*, 31(6):385–93.

Perkins, D. O., Gu, H., Boteva, K., and Lieberman, J. A. (2005). Relationship Between Duration of Untreated Psychosis and Outcome in First-Episode Schizophrenia: A Critical Review and Meta-Analysis. *American Journal of Psychiatry*, 162(10):1785–1804.

Perkins, D. O., Jeffries, C. D., Addington, J., Bearden, C. E., Cadenhead, K. S., Cannon, T. D., Cornblatt, B. A., Mathalon, D. H., McGlashan, T. H., Seidman, L. J., Tsuang, M. T., Walker, E. F., Woods, S. W., and Heinssen, R. (2015). Towards a Psychosis Risk Blood Diagnostic for Persons Experiencing High-Risk Symptoms: Preliminary Results From the NAPLS Project. *Schizophrenia Bulletin*, 41(2):419–428.

Pirooznia, M., Wang, T., Avramopoulos, D., Potash, J. B., Zandi, P. P., and Goes, F. S. (2016). High-throughput sequencing of the synaptome in major depressive disorder. 21(5):650–655.

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709.

Prasad, K. M., Watson, A. M. M., Dickerson, F. B., Yolken, R. H., and Nimgaonkar, V. L. (2012). Exposure to Herpes Simplex Virus Type 1 and Cognitive Impairments in Individuals With Schizophrenia. *Schizophrenia Bulletin*, 38(6):1137–1148.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Psychosis Endophenotypes International Consortium, P. E. I., Wellcome Trust Case-Control Consortium 2, W. T. C.-C. C., Bramon, E., Pirinen, M., Strange, A., Lin, K., Freeman, C., Bellenguez, C., Su, Z., Band, G., Pearson, R., Vukcevic, D., Langford, C., Deloukas, P., Hunt, S., Gray, E., Dronov, S., Potter, S. C., Tashakkori-Ghanbaria, A., Edkins, S., Bumpstead, S. J., Arranz, M. J., Bakker, S., Bender, S., Bruggeman, R., Cahn, W., Chandler, D., Collier, D. A., Crespo-Facorro, B., Dazzan, P., de Haan, L., Di Forti, M., Dragović, M., Giegling, I., Hall, J., Iyegbe, C., Jablensky, A., Kahn, R. S., Kalaydjieva, L., Kravariti, E., Lawrie, S., Linszen, D. H., Mata, I., McDonald, C., McIntosh, A., Myin-Germeys, I., Ophoff, R. A., Pariante, C. M., Paunio, T., Picchioni, M., Psychiatric Genomics Consortium, Ripke, S., Rujescu, D., Sauer, H., Shaikh, M., Sussmann, J., Suvisaari, J., Tosato, S., Toulopoulou, T., Van Os, J., Walshe, M., Weisbrod, M., Whalley, H., Wiersma, D., Blackwell, J. M., Brown, M. A., Casas, J. P., Corvin, A., Duncanson, A., Jankowski, J. A. Z., Markus, H. S., Mathew, C. G., Palmer, C. N. A., Plomin, R., Rautanen, A., Sawcer, S. J., Trembath, R. C., Wood, N. W., Barroso, I., Peltonen, L., Lewis, C. M., Murray, R. M., Donnelly, P., Powell, J., and Spencer, C. C. A. (2014). A genome-wide association analysis of a broad psychosis phenotype identifies three loci for further investigation. *Biological psychiatry*, 75(5):386–97.

Purcell, S., Moran, J., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S., Kähler, A., Duncan, L., Stahl, E., Genovese, G., Fernández, E., Collins, M., Komiyama, N., Choudhary, J., Magnusson, P., Banks, E., Shakir, K., Garimella, K., Fennell, T., Depristo, M., Grant, S., Haggarty, S., Gabriel, S., Scolnick, E., Lander, E., Hultman, C., Sullivan, P., McCarroll, S., and Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487).

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Ruderfer, D. M., McQuillin, A., Morris, D. W., Oĝdushlaine, C. T., Corvin, A., Holmans, P. A., Oĝdonovan, M. C., MacGregor, S., Gurling, H., Blackwood, D. H. R., Craddock, N. J., Gill, M., Hultman, C. M., Kirov, G. K., Lichtenstein, P., Muir, W. J., Owen, M. J., Pato, C. N., Scolnick, E. M., St Clair, D., Williams, N. M., Georgieva, L., Nikolov, I., Norton, N., Williams, H., Toncheva, D., Milanova, V., Thelander, E. F., O'Dushlaine, C. T., Kenny, E., Quinn, E. M., Choudhury, K., Datta, S., Pimm, J., Thirumalai, S., Puri, V., Krasucki, R., Lawrence, J., Quested, D., Bass, N., Crombie, C., Fraser, G., Leh Kuan, S., Walker, N., McGhee, K. A., Pickard, B., Malloy, P., MacLean, A. W., Van Beck, M., Pato, M. T., Medeiros, H., Middleton, F., Carvalho, C., Morley, C., Fanous, A., Conti, D., Knowles, J. A., Paz Ferreira, C., MacEdo, A., Helena Azevedo, M., Kirby, A. N., Ferreira, M. A. R., Daly, M. J., Chambert, K., Kuruvilla, F., Gabriel, S. B., Ardlie, K., Moran, J. L., and Sklar, P. (2009a). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. 460(7256):748–752.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. (2009b). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–52.

Rees, E., Walters, J. T. R., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L., Mahoney-Davies, G., Legge, S. E., Moran, J. L., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., and Kirov, G. (2014). Analysis of copy number variations at 15 schizophrenia-associated loci. *The British journal of psychiatry : the journal of mental science*, 204(2):108–14.

Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers.

Rice, G. I., Kasher, P. R., Forte, G. M. A., Mannion, N. M., Greenwood, S. M., Szynkiewicz, M., Dickerson, J. E., Bhaskar, S. S., Zampini, M., Briggs, T. A., Jenkinson, E. M., Bacino, C. A., Battini, R., Bertini, E., Brogan, P. A., Brueton, L. A., Carpanelli, M., De Laet, C., de Lonlay, P., del Toro, M., Desguerre, I., Fazzi, E., Garcia-Cazorla, A., Heiberg, A., Kawaguchi, M., Kumar, R., Lin, J.-P. S.-M., Lourenco, C. M., Male, A. M., Marques, W., Mignot, C., Olivieri, I., Orcesi, S., Prabhakar, P., Rasmussen, M., Robinson, R. A., Rozenberg, F., Schmidt, J. L., Steindl, K., Tan, T. Y., van der Merwe, W. G., Vanderver, A., Vassallo, G., Wakeling, E. L., Wassmer, E., Whittaker, E., Livingston, J. H., Lebon, P., Suzuki, T., McLaughlin, P. J., Keegan, L. P., O'Connell, M. A., Lovell, S. C., and Crow, Y. J. (2012). Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. *Nature Genetics*, 44(11):1243–1248.

Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package.

Rijsdijk, F. V. and Sham, P. C. (2002). Analytic approaches to twin data using structural equation models. *Briefings in bioinformatics*, 3(2):119–33.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. a., Lee, P., Bulik-Sullivan, B., Collier, D. a., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. a., Begemann, M., Belliveau Jr, R. a., Bene, J., Bergen, S. E., Bevilacqua, E., Bigdeli, T. B., Black, D. W., Bruggeman, R., Buccola, N. G., Buckner, R. L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R. M., Carr, V. J., Carrera, N., Catts, S. V., Chambert, K. D., Chan, R. C. K., Chen, R. Y. L., Chen, E. Y. H., Cheng, W., Cheung, E. F. C., Ann Chong, S., Robert Cloninger, C., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J. J., Curtis, D., Davidson, M., Davis, K. L., Degenhardt, F., Del Favero, J., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A. H., Farrell, M. S., Frank, J., Franke, L., Freedman, R., Freimer, N. B., Friedl, M., Friedman, J. I., Fromer, M., Genovese, G., Georgieva, L., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J. I., Golimbet, V., Gopal, S., Gratten, J., de Haan, L., Hammer, C., Hamshere, M. L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A. M., Henskens, F. a., Herms, S., Hirschhorn, J. N., Hoffmann, P., Hofman, A., Hollegaard, M. V., Hougaard, D. M., Ikeda, M., Joa, I., Julià, A., Kahn, R. S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M. C., Kennedy, J. L., Khrunin, A., Kim, Y., Klovins, J., Knowles, J. a., Konte, B., Kucinskas, V., Ausrele Kucinskiene, Z., Kuzelova-Ptackova, H., Kähler, A. K., Laurent, C., Lee Chee Keong, J., Hong Lee, S., Legge, S. E., Lerer, B., Li, M., Li, T., Liang, K.-Y., Lieberman, J., Limborska, S., Loughland, C. M., Lubinski, J., Lönnqvist, J., Macek Jr, M., Magnusson, P. K. E., Maher, B. S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R. W., McDonald, C., McIntosh, A. M., Meier, S., Meijer, C. J., Melegh, B., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mokrab, Y., Morris, D. W., Mors, O., Murphy, K. C., Murray, R. M., Myin-Germeys, I., Müller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D. a., Nestadt, G., Nicodemus, K. K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F. A., Oh, S.-Y., Olincy, A., Olsen, L., Van Os, J., Endophenotypes International Consortium, P., Pantelis, C., Papadimitriou, G. N., Papiol, S., Parkhomenko, E., Pato, M. T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D. O., Pietiläinen, O., Pimm, J., Pocklington, A. J., Powell, J., Price, A., Pulver, A. E., Purcell, S. M., Quested, D., Rasmussen, H. B., Reichenberg, A., Reimers, M. a., Richards, A. L., Roffman, J. L., Roussos, P., Ruderfer, D. M., Salomaa, V., Sanders, A. R., Schall, U., Schubert, C. R., Schulze, T. G., Schwab, S. G., Scolnick, E. M., Scott, R. J., Seidman, L. J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J. M., Sim, K., Slominsky, P., Smoller, J. W., So, H.-C., Spencer, C. a., Stahl, E. a., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R. E., Strengman, E., Strohmaier, J., Scott Stroup, T., Subramaniam, M., Suvisaari, J., Svrakic, D. M., Szatkiewicz, J. P., Söderman, E., Thirumalai, S., Toncheva, D., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B. T., Weiser, M., Wildenauer, D. B., Williams, N. M., Williams, S., Witt, S. H., Wolen, A. R., Wong, E. H. M., Wormley, B. K., Simon Xi, H., Zai, C. C., Zheng, X., Zimprich, F., Wray, N. R., Stefansson, K., Visscher, P. M., Trust Case-Control Consortium, W., Adolfsson, R., Andreassen, O. a., Blackwood, D. H. R., Bramon, E., Buxbaum, J. D., Børglum, A. D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C. M., Iwata, N., Jablensky, A. V., Jönsson, E. G., Kendler, K. S., Kirov, G., Knight, J., Lencz, T., Levinson, D. F., Li, Q. S., Liu, J., Malhotra, A. K., McCarroll, S. a., McQuillin, A., Moran, J. L., Mortensen, P. B., Mowry, B. J., Nöthen, M. M., Ophoff, R. a., Owen, M. J., Palotie, A., Pato, C. N., Petryshen, T. L., Posthuma, D.,

Rietschel, M., Riley, B. P., Rujescu, D., Sham, P. C., Sklar, P., St Clair, D., Weinberger, D. R., Wendland, J. R., Werge, T., Daly, M. J., Sullivan, P. F., and O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*.

Ripley, B. D. (1996). 'Pattern Recognition and Neural Networks'. *Press*, pages 0–521.

Robinson, J. E., Paluch, J., Dickman, D. K., and Joiner, W. J. (2016). ADAR-mediated RNA editing suppresses sleep by acting as a brake on glutamatergic synaptic plasticity. *Nature Communications*, 7:10512.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. 65(6):386–408.

Rucker, J., Newman, S., Gray, J., Gunasinghe, C., Broadbent, M., Brittain, P., Baggaley, M., Denis, M., Turp, J., Stewart, R., Lovestone, S., Schumann, G., Farmer, A., and McGuffin, P. (2011). OPCRIT+: an electronic system for psychiatric diagnosis and data collection in clinical and research settings. *The British Journal of Psychiatry*, 199(2):151–155.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pedro Pereira, R., Pilicheva, E., Rung, J., Sharma, A., Tang, Y. A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., and Sarkans, U. (2013). ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic acids research*, 41(Database issue):987–90.

Saetre, P., Emilsson, L., Axelsson, E., Kreuger, J., Lindholm, E., and Jazin, E. (2007). Inflammation-related genes up-regulated in schizophrenia brains. *BMC psychiatry*, 7:46.

Salameh, A., Lee, A. K., Cardó-Vila, M., Nunes, D. N., Efstathiou, E., Staquicini, F. I., Dobroff, A. S., Marchiò, S., Navone, N. M., Hosoya, H., Lauer, R. C., Wen, S., Salmeron, C. C., Hoang, A., Newsham, I., Lima, L. A., Carraro, D. M., Oliviero, S., Kolonin, M. G., Sidman, R. L., Do, K.-A., Troncoso, P., Logothetis, C. J., Brentani, R. R., Calin, G. A., Cavenee, W. K., Dias-Neto, E., Pasqualini, R., and Arap, W. (2015). PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proceedings of the National Academy of Sciences*, 112(27):8403–8408.

Sanders, A. R., Duan, J., Levinson, D. F., Shi, J., He, D., Hou, C., Burrell, G. J., Rice, J. P., Nertney, D. A., Olincy, A., Rozic, P., Vinogradov, S., Buccola, N. G., Mowry, B. J., Freedman, R., Amin, F., Black, D. W., Silverman, J. M., Byerley, W. F., Crowe, R. R., Cloninger, C. R., Martinez, M., and Gejman, P. V. (2008). No Significant Association of 14 Candidate Genes With Schizophrenia in a Large European Ancestry Sample: Implications for Psychiatric Genetics. *American Journal of Psychiatry*, 165(4):497–506.

Sanders, A. R., Göring, H. H. H., Duan, J., Drigalenko, E. I., Moy, W., Freda, J., He, D., Shi, J., and Gejman, P. V. (2013). Transcriptome study of differential expression in schizophrenia. *Human molecular genetics*, 22(24):5001–14.

Sartor, C. E., McCutcheon, V. V., Pommer, N. E., Nelson, E. C., Grant, J. D., Duncan, A. E., Waldron, M., Bucholz, K. K., Madden, P. A. F., and Heath, A. C. (2011). Common genetic and environmental contributions to post-traumatic stress disorder and alcohol dependence in young women. *Psychological Medicine*, 41(07):1497–1505.

Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., Eils, R., Weith, A., Mennerich, D., and Quast, K. (2010). Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC genomics*, 11:349.

Schneider, M., Debbané, M., Bassett, A. S., Chow, E. W. C., Fung, W. L. A., Van Den Bree, M. B. M., Owen, M., Murphy, K. C., Niarchou, M., Kates, W. R., Antshel, K. M., Fremont, W., McDonald-McGinn, D. M., Gur, R. E., Zackai, E. H., Vorstman, J., Duijff, S. N., Klaassen, P. W. J., Swillen, A., Gothelf, D., Green, T., Weizman, A., Van Amelsvoort, T., Evers, L., Boot, E., Shashi, V., Hooper, S. R., Bearden, C. E., Jalbrzikowski, M., Armando, M., Vicari, S., Murphy, D. G., Ousley, O., Campbell, L. E., Simon, T. J., and Eliez, S. (2014). Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: Results from the international consortium on brain and behavior in 22q11.2 deletion syndrome.

Schönemann, P. H. (1997). On models and muddles of heritability. *Genetica*, 99(2/3):97–108.

Sekar, A., Bialas, A. R., Rivera, H. d., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Doren, V. V., Genovese, G., Rose, S. A., Handsaker, R. E., Consortium, S. W. G. o. t. P. G., Daly, M. J., Carroll, M. C., Stevens, B., and McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177.

Shoja Shafti, S. and Gilanipoor, M. (2014). A Comparative Study between Olanzapine and Risperidone in the Management of Schizophrenia. *Schizophrenia research and treatment*, 2014:307202.

Shprintzen, R. J. (2008). Velo-cardio-facial syndrome: 30 Years of study.

Siemens, H. (1924). *Die zwillingspathologie ihre bedeutung, ihre methodik, ihre bisherigen ergebnisse,*. J. Springer, Berlin.

Sirabella, R., Secondo, A., Pannaccione, A., Scorziello, A., Valsecchi, V., Adornetto, A., Bilo, L., Di Renzo, G., and Annunziato, L. (2009). Anoxia-induced NF-kB-dependent upregulation of NCX1 contributes to Ca 2+ refilling into endoplasmic reticulum in cortical neurons. 40(3):922–929.

Smoller, J. W. and Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics*, 123C(1):48–58.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing di erential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, Vol. 3(No. 1):Article 3.

Stefansson, H., Ophoff, R. a., Steinberg, S., Andreassen, O. a., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O. P. H., Mors, O., Mortensen, P. B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Børglum, A. D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Böttcher, Y., Olesen, J., Breuer, R., Möller, H.-J., Giegling, I., Rasmussen, H. B., Timm, S., Mattheisen, M., Bitter, I., Réthelyi, J. M., Magnusdottir, B. B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J. R., Haraldsson, M., Fossdal, R., Thorgeirsson, T. E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman,

E., Kiemeney, L. a., Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Ettinger, U., Picchioni, M., Walker, N., Toulopoulou, T., Need, A. C., Ge, D., Yoon, J. L., Shianna, K. V., Freimer, N. B., Cantor, R. M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jönsson, E. G., Terenius, L., Agartz, I., Petursson, H., Nöthen, M. M., Rietschel, M., Matthews, P. M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D. B., Stefansson, K., and Collier, D. a. (2009). Common variants conferring risk of schizophrenia. *Nature*, 460(7256):744–7.

Stieglitz, B., Morris-Davies, A. C., Koliopoulos, M. G., Christodoulou, E., and Rittinger, K. (2012). LUBAC synthesizes linear ubiquitin chains via a thioester intermediate. *EMBO reports*, 13(9):840–6.

Sullivan, K. E. (2004). The clinical, immunological, and molecular spectrum of chromosome 22q11.2 deletion syndrome and DiGeorge syndrome.

Sullivan, P. F. (2005). The genetics of schizophrenia. *PLoS medicine*, 2(7):e212.

Sullivan, P. F., Kendler, K. S., and Neale, M. C. (2003). Schizophrenia as a Complex Trait. 60.

Sullivan, P. F., Neale, M. C., and Kendler, K. S. (2000). Genetic Epidemiology of Major Depression: Review and Meta-Analysis. *American Journal of Psychiatry*, 157(10):1552–1562.

Takahashi, M., Hayashi, H., Watanabe, Y., Sawamura, K., Fukui, N., Watanabe, J., Kitajima, T., Yamanouchi, Y., Iwata, N., Mizukami, K., Hori, T., Shimoda, K., Ujike, H., Ozaki, N., Iijima, K., Takemura, K., Aoshima, H., and Someya, T. (2010). Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures. *Schizophrenia Research*, 119(1-3):210–218.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tokunaga, F. and Iwai, K. (2012). LUBAC, a novel ubiquitin ligase for linear ubiquitination, is crucial for inflammation and immune responses. *Microbes and infection / Institut Pasteur*, 14(7-8):563–72.

Turkheimer, E. (2016). Weak Genetic Explanation 20 Years Later. *Perspectives on Psychological Science*, 11(1):24–28.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., and Gottesman, I. I. (2003). Socioeconomic Status Modifies Heritability of IQ in Young Children. *Psychological Science*, 14(6):623–628.

Vassos, E., Di Forti, M., Coleman, J., Iyegbe, C., Prata, D., Euesden, J., O'Reilly, P., Curtis, C., Kolliakou, A., Patel, H., Newhouse, S., Traylor, M., Ajnakina, O., Mondelli, V., Marques, T., Gardner-Sood, P., Aitchison, K., Powell, J., Atakan, Z., Greenwood, K., Smith, S., Ismail, K., Pariante, C., Gaughran, F., Dazzan, P., Markus, H., David, A., Lewis, C., Murray, R., and Breen, G. (2017). An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. *Biological Psychiatry*, 81(6).

Vicente-Sánchez, A., Sánchez-Blázquez, P., Rodríguez-Muñoz, M., and Garzón, J. (2013). HINT1 protein cooperates with cannabinoid 1 receptor to negatively regulate glutamate NMDA receptor activity. *Molecular Brain*, 6(1):42.

Vilhjálmsson, B. J. and Nordborg, M. (2012). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2.

Wang, Q., Li, X., Qi, R., and Billiar, T. (2017). RNA Editing, ADAR1, and the Innate Immune Response. *Genes*, 8(1).

Wei, C., Zhou, J., Huang, X., and Li, M. (2008). Effects of psychological stress on serum iron and erythropoiesis. *International Journal of Hematology*, 88(1):52–56.

Weihs, C., Ligges, U., Luebke, K., and Nils, R. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

Weinstock, M. (2008). The long-term behavioural consequences of prenatal stress. *Neuroscience & Biobehavioral Reviews*, 32(6):1073–1086.

Wong, B. X. and Duce, J. A. (2014). The iron regulatory capability of the major protein participants in prevalent neurodegenerative disorders. *Frontiers in Pharmacology*, 5:81.

Wu, J. Q., Green, M. J., Gardiner, E. J., Tooney, P. A., Scott, R. J., Carr, V. J., and Cairns, M. J. (2016). Altered neural signaling and immune pathways in peripheral blood mononuclear cells of schizophrenia patients with cognitive impairment: A transcriptome analysis. *Brain, Behavior, and Immunity*, 53:194–206.

Xie, Y. (2010). MBCB: MBCB (Model-based Background Correction for Beadarray).

Xie, Y., Wang, X., and Story, M. (2009). Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*, 25(6):751–757.

Yang, W., Thompson, J. W., Wang, Z., Wang, L., Sheng, H., Foster, M. W., Moseley, M. A., and Paschen, W. (2012). Analysis of oxygen/glucose-deprivation-induced changes in SUMO3 conjugation using SILAC-based quantitative proteomics. *Journal of proteome research*, 11(2):1108–17.

Yang, W., Wang, Q., Kanes, S. J., Murray, J. M., and Nishikura, K. (2004). Altered RNA editing of serotonin 5-HT2C receptor induced by interferon: implications for depression associated with cytokine therapy. *Molecular Brain Research*, 124(1):70–78.

Yolken, R. (2004). Viruses and schizophrenia: a focus on herpes simplex virus. *Herpes : the journal of the IHMF*, 11 Suppl 2:83A–88A.

Zammit, S., Spurlock, G., Williams, H., Norton, N., Williams, N., O'Donovan, M. C., and Owen, M. J. (2007). Genotype effects of CHRNA7, CNRI and COMT in schizophrenia: Interactions with tobacco and cannabis use. 191(NOV.):402–407.

Zandi, M. S., Irani, S. R., Lang, B., Waters, P., Jones, P. B., McKenna, P., Coles, A. J., Vincent, A., and Lennox, B. R. (2011). Disease-relevant autoantibodies in first episode schizophrenia. *Journal of Neurology*, 258(4):686–688.

Appendix A

Supplementary Material: Chapter 3

Full List of differentially Expressed probes.

Table A.1 Top differentially expressed probes (Complete)

Gene	logFC	p-value	q-value	CHR	Definition
Up-regulated					
SUMO3	0.1	2.37E-08	0.0000736	21	Homo sapiens SMT3 suppressor of mif two 3 homolog 3 (S. cerevisiae) (SUMO3), mRNA.
CAMP	0.63	3.11E-08	0.0000736	3	Homo sapiens cathelicidin antimicrobial peptide (CAMP), mRNA.
DEFA1B	0.91	0.000000209	0.000206	8	Homo sapiens defensin, alpha 1B (DEFA1B), mRNA.
DEFA1	0.82	0.000000323	0.000206	8	Homo sapiens defensin, alpha 1 (DEFA1), mRNA.
DEFA3	0.9	0.000000353	0.000206	8	Homo sapiens defensin, alpha 3, neutrophil-specific (DEFA3), mRNA.
IDH1	0.13	0.000000375	0.000206	2	Homo sapiens isocitrate dehydrogenase 1 (NADP+), soluble (IDH1), mRNA.
TMEM170B	0.29	0.000000391	0.000206	6	Homo sapiens transmembrane protein 170B (TMEM170B), mRNA.
LDHA	0.1	0.000000562	0.000239	11	Homo sapiens lactate dehydrogenase A (LDHA), transcript variant 2, mRNA.
LCN2	0.58	0.000000607	0.000239	9	Homo sapiens lipocalin 2 (LCN2), mRNA.
C9ORF72	0.23	0.000000871	0.000276	9	Homo sapiens chromosome 9 open reading frame 72 (C9orf72), transcript variant 1, mRNA.
LYPLAL1	0.17	0.000000905	0.000276	1	Homo sapiens lysophospholipase-like 1 (LYPLAL1), mRNA.
TFRC	0.17	0.00000104	0.000276	3	Homo sapiens transferrin receptor (p90, CD71) (TFRC), mRNA.
S100A8	0.35	0.00000105	0.000276	1	Homo sapiens S100 calcium binding protein A8 (S100A8), mRNA.
PCMT1	0.13	0.00000114	0.000284	6	Homo sapiens protein-L-isoaspartate (D-aspartate) O-methyltransferase (PCMT1), mRNA.
BNIP2	0.17	0.0000013	0.000307	15	Homo sapiens BCL2/adenovirus E1B 19kDa interacting protein 2 (BNIP2), mRNA.
SLC30A9	0.17	0.00000137	0.000307	4	Homo sapiens solute carrier family 30 (zinc transporter), member 9 (SLC30A9), mRNA.
SLC44A1	0.17	0.00000169	0.000348	9	Homo sapiens solute carrier family 44, member 1 (SLC44A1), mRNA.
TCEB1	0.13	0.00000191	0.000362	8	Homo sapiens transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C) (TCEB1), mRNA.
VAMP7	0.14	0.00000191	0.000362	XY	Homo sapiens vesicle-associated membrane protein 7 (VAMP7), mRNA.
GLRX	0.22	0.00000224	0.000367	5	Homo sapiens glutaredoxin (thioltransferase) (GLRX), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
CLNS1A	0.17	0.00000237	0.000367	11	Homo sapiens chloride channel, nucleotide-sensitive, 1A (CLNS1A), mRNA.
FAM96A	0.23	0.00000239	0.000367	15	Homo sapiens family with sequence similarity 96, member A (FAM96A), transcript variant 1, mRNA.
H2AFZ	0.1	0.00000242	0.000367	4	Homo sapiens H2A histone family, member Z (H2AFZ), mRNA.
SENP7	0.25	0.00000271	0.000367	3	Homo sapiens SUMO1/sentrin specific peptidase 7 (SENP7), transcript variant 2, mRNA.
COX7A2L	0.18	0.00000271	0.000367	2	Homo sapiens cytochrome c oxidase subunit VIIa polypeptide 2 like (COX7A2L), nuclear gene encoding mitochondrial protein, mRNA.
C14ORF100	0.14	0.00000286	0.000367	14	Homo sapiens chromosome 14 open reading frame 100 (C14orf100), mRNA.
TAF7	0.19	0.00000288	0.000367	5	Homo sapiens TAF7 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 55kDa (TAF7), mRNA.
PSMC2	0.14	0.00000299	0.000367	7	Homo sapiens proteasome (prosome, macropain) 26S subunit, ATPase, 2 (PSMC2), mRNA.
S100A12	0.3	0.00000303	0.000367	1	Homo sapiens S100 calcium binding protein A12 (S100A12), mRNA.
MED28	0.12	0.00000329	0.00038	4	Homo sapiens mediator complex subunit 28 (MED28), mRNA.
ARL6IP5	0.13	0.0000035	0.000394	3	Homo sapiens ADP-ribosylation-like factor 6 interacting protein 5 (ARL6IP5), mRNA.
IFNGR1	0.15	0.0000043	0.000433	6	Homo sapiens interferon gamma receptor 1 (IFNGR1), mRNA.
SRP9	0.19	0.00000457	0.000434	1	Homo sapiens signal recognition particle 9kDa (SRP9), mRNA.
PRDX3	0.12	0.00000459	0.000434	10	Homo sapiens peroxiredoxin 3 (PRDX3), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
ATG3	0.14	0.00000495	0.000436	3	Homo sapiens ATG3 autophagy related 3 homolog (S. cerevisiae) (ATG3), mRNA.
CRLS1	0.2	0.00000507	0.000436	20	Homo sapiens cardiolipin synthase 1 (CRLS1), mRNA.
CCPG1	0.16	0.00000543	0.000445	15	Homo sapiens cell cycle progression 1 (CCPG1), transcript variant 2, mRNA.
CLDND1	0.18	0.00000545	0.000445	3	Homo sapiens claudin domain containing 1 (CLDND1), transcript variant 1, mRNA.
ANXA3	0.34	0.00000571	0.000458	4	Homo sapiens annexin A3 (ANXA3), mRNA.
TM2D1	0.11	0.00000633	0.000466	1	Homo sapiens TM2 domain containing 1 (TM2D1), mRNA.
MAP2K1IP1	0.2	0.00000657	0.000466	4	Homo sapiens mitogen-activated protein kinase kinase 1 interacting protein 1 (MAP2K1IP1), mRNA.
COX7A2	0.21	0.0000067	0.000466	6	Homo sapiens cytochrome c oxidase subunit VIIa polypeptide 2 (liver) (COX7A2), mRNA.
FAM45A	0.13	0.00000696	0.000477	10	Homo sapiens family with sequence similarity 45, member A (FAM45A), mRNA.
ATP5C1	0.23	0.00000764	0.000509	10	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1 (ATP5C1), nuclear gene encoding mitochondrial protein, transcript variant 2, mRNA.
CDC40	0.11	0.000008	0.00052	6	Homo sapiens cell division cycle 40 homolog (S. cerevisiae) (CDC40), mRNA.
VBP1	0.21	0.00000802	0.00052	X	Homo sapiens von Hippel-Lindau binding protein 1 (VBP1), mRNA.
PHF5A	0.19	0.00000872	0.000529	22	Homo sapiens PHD finger protein 5A (PHF5A), mRNA.
GNG10	0.3	0.00000901	0.000533	9	Homo sapiens guanine nucleotide binding protein (G protein), gamma 10 (GNG10), mRNA.
PSMD10	0.15	0.00000935	0.000533	X	Homo sapiens proteasome (prosome, macropain) 26S subunit, non-ATPase, 10 (PSMD10), transcript variant 1, mRNA.
HEXB	0.1	0.00000953	0.000533	5	Homo sapiens hexosaminidase B (beta polypeptide) (HEXB), mRNA.
YPEL5	0.13	0.00000973	0.000533	2	Homo sapiens yippee-like 5 (Drosophila) (YPEL5), mRNA.
WDR61	0.13	0.00000973	0.000533	15	Homo sapiens WD repeat domain 61 (WDR61), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
COX17	0.13	0.00001	0.000533	3	Homo sapiens COX17 cytochrome c oxidase assembly homolog (S. cerevisiae) (COX17), nuclear gene encoding mitochondrial protein, mRNA.
UQCRQ	0.26	0.0000102	0.000533	5	Homo sapiens ubiquinol-cytochrome c reductase, complex III subunit VII, 9.5kDa (UQCRQ), nuclear gene encoding mitochondrial protein, mRNA.
PHF20L1	0.16	0.0000104	0.000533	8	Homo sapiens PHD finger protein 20-like 1 (PHF20L1), transcript variant 1, mRNA.
KLHDC2	0.15	0.0000114	0.000558	14	Homo sapiens kelch domain containing 2 (KLHDC2), mRNA.
CD24	0.27	0.0000117	0.000558	Y	Homo sapiens CD24 molecule (CD24), mRNA.
TANK	0.12	0.0000117	0.000558	2	Homo sapiens TRAF family member-associated NFKB activator (TANK), transcript variant 1, mRNA.
LYST	0.13	0.0000118	0.000558	1	Homo sapiens lysosomal trafficking regulator (LYST), mRNA.
FBXL5	0.12	0.0000122	0.000563	4	Homo sapiens F-box and leucine-rich repeat protein 5 (FBXL5), transcript variant 1, mRNA.
PIGY	0.17	0.0000124	0.000565	4	Homo sapiens phosphatidylinositol glycan anchor biosynthesis, class Y (PIGY), transcript variant 2, mRNA.
ENY2	0.19	0.0000131	0.000579	8	Homo sapiens enhancer of yellow 2 homolog (Drosophila) (ENY2), mRNA.
TBK1	0.15	0.0000132	0.000579	12	Homo sapiens TANK-binding kinase 1 (TBK1), mRNA.
ORMDL1	0.14	0.0000137	0.000587	2	Homo sapiens ORM1-like 1 (S. cerevisiae) (ORMDL1), mRNA.
ASNSD1	0.19	0.0000141	0.000597	2	Homo sapiens asparagine synthetase domain containing 1 (ASNSD1), mRNA.
CYB5R4	0.13	0.0000147	0.000617	6	Homo sapiens cytochrome b5 reductase 4 (CYB5R4), mRNA.
TXNDC17	0.19	0.0000152	0.000632	17	Homo sapiens thioredoxin domain containing 17 (TXNDC17), mRNA.
SRGN	0.12	0.0000163	0.000658	10	Homo sapiens serglycin (SRGN), mRNA.
TMX1	0.24	0.0000165	0.000658	14	Homo sapiens thioredoxin-related transmembrane protein 1 (TMX1), mRNA.
TXN	0.17	0.0000167	0.000659	9	Homo sapiens thioredoxin (TXN), mRNA.
RNF7	0.14	0.0000174	0.000676	3	Homo sapiens ring finger protein 7 (RNF7), transcript variant 3, mRNA.
COMMD3	0.16	0.0000189	0.000696	10	Homo sapiens COMM domain containing 3 (COMMD3), mRNA.
KIAA1600	0.18	0.0000193	0.000696	10	Homo sapiens KIAA1600 (KIAA1600), mRNA.
UBL3	0.17	0.0000197	0.000696	13	Homo sapiens ubiquitin-like 3 (UBL3), mRNA.
MTMR6	0.17	0.0000199	0.000696	13	Homo sapiens myotubularin related protein 6 (MTMR6), mRNA.
OSBPL8	0.22	0.00002	0.000696	12	Homo sapiens oxysterol binding protein-like 8 (OSBPL8), transcript variant 2, mRNA.
MARCH7	0.15	0.0000205	0.000703	2	Homo sapiens membrane-associated ring finger (C3HC4) 7 (MARCH7), mRNA.
CNIH4	0.26	0.0000211	0.000706	1	Homo sapiens cornichon homolog 4 (Drosophila) (CNIH4), mRNA.
SLC35A1	0.18	0.0000212	0.000706	6	Homo sapiens solute carrier family 35 (CMP-sialic acid transporter), member A1 (SLC35A1), mRNA.
COPS5	0.13	0.0000212	0.000706	8	Homo sapiens COP9 constitutive photomorphogenic homolog subunit 5 (Arabidopsis) (COPS5), mRNA.
TMED7	0.16	0.0000214	0.000706	5	Homo sapiens transmembrane emp24 protein transport domain containing 7 (TMED7), mRNA.
KBTBD11	0.15	0.0000216	0.000708	8	Homo sapiens kelch repeat and BTB (POZ) domain containing 11 (KBTBD11), mRNA.
MRPL3	0.2	0.0000221	0.000717	3	Homo sapiens mitochondrial ribosomal protein L3 (MRPL3), nuclear gene encoding mitochondrial protein, mRNA.
MRPL33	0.11	0.0000223	0.000717	2	Homo sapiens mitochondrial ribosomal protein L33 (MRPL33), nuclear gene encoding mitochondrial protein, transcript variant 2, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
PPP3CB	0.1	0.000023	0.000726	10	Homo sapiens protein phosphatase 3 (formerly 2B), catalytic subunit, beta isoform (PPP3CB), mRNA.
AGL	0.23	0.0000241	0.000726	1	Homo sapiens amylo-1, 6-glucosidase, 4-alpha-glucanotransferase (AGL), transcript variant 5, mRNA.
TOMM20	0.11	0.0000245	0.000726	1	Homo sapiens translocase of outer mitochondrial membrane 20 homolog (yeast) (TOMM20), nuclear gene encoding mitochondrial protein, mRNA.
LY96	0.42	0.0000249	0.000726	8	Homo sapiens lymphocyte antigen 96 (LY96), mRNA.
CNOT8	0.12	0.000025	0.000726	5	Homo sapiens CCR4-NOT transcription complex, subunit 8 (CNOT8), mRNA.
MYL6	0.12	0.0000251	0.000726	12	Homo sapiens myosin, light chain 6, alkali, smooth muscle and non-muscle (MYL6), transcript variant 1, mRNA.
RPSA	0.15	0.0000252	0.000726	3	Homo sapiens ribosomal protein SA (RPSA), transcript variant 1, mRNA.
RCBTB2	0.14	0.0000252	0.000726	13	Homo sapiens regulator of chromosome condensation (RCC1) and BTB (POZ) domain containing protein 2 (RCBTB2), mRNA.
AGTPBP1	0.11	0.0000261	0.000738	9	Homo sapiens ATP/GTP binding protein 1 (AGTPBP1), mRNA.
UBLCP1	0.11	0.0000263	0.000742	5	Homo sapiens ubiquitin-like domain containing CTD phosphatase 1 (UBLCP1), mRNA.
MRPL32	0.19	0.0000271	0.000754	7	Homo sapiens mitochondrial ribosomal protein L32 (MRPL32), nuclear gene encoding mitochondrial protein, mRNA.
TMEM14B	0.15	0.0000278	0.000759	6	Homo sapiens transmembrane protein 14B (TMEM14B), mRNA.
AZIN1	0.14	0.0000282	0.000763	8	Homo sapiens antizyme inhibitor 1 (AZIN1), transcript variant 1, mRNA.
SF3B14	0.2	0.0000288	0.000771	2	Homo sapiens splicing factor 3B, 14 kDa subunit (SF3B14), mRNA.
TPK1	0.11	0.0000297	0.000777	7	Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant 2, mRNA.
STK17B	0.16	0.0000306	0.000784	2	Homo sapiens serine/threonine kinase 17b (STK17B), mRNA.
CKLF	0.14	0.0000306	0.000784	16	Homo sapiens chemokine-like factor (CKLF), transcript variant 5, mRNA.
ANGEL2	0.12	0.0000309	0.000785	1	Homo sapiens angel homolog 2 (Drosophila) (ANGEL2), mRNA.
SUMO1P3	0.1	0.0000325	0.000818	1	Homo sapiens SUMO1 pseudogene 3 (SUMO1P3), non-coding RNA.
GGPS1	0.1	0.0000332	0.000826	1	Homo sapiens geranylgeranyl diphosphate synthase 1 (GGPS1), transcript variant 2, mRNA.
HIGD1A	0.18	0.0000334	0.000826	3	Homo sapiens HIG1 hypoxia inducible domain family, member 1A (HIGD1A), transcript variant 1, mRNA.
UBE2E1	0.12	0.0000336	0.000829	3	Homo sapiens ubiquitin-conjugating enzyme E2E 1 (UBC4/5 homolog, yeast) (UBE2E1), transcript variant 2, mRNA.
ANKIB1	0.15	0.0000361	0.000846	7	Homo sapiens ankyrin repeat and IBR domain containing 1 (ANKIB1), mRNA.
SRP19	0.14	0.0000362	0.000846	5	Homo sapiens signal recognition particle 19kDa (SRP19), mRNA.
DUSP12	0.12	0.0000362	0.000846	1	Homo sapiens dual specificity phosphatase 12 (DUSP12), mRNA.
DPM1	0.19	0.0000372	0.000846	20	Homo sapiens dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit (DPM1), mRNA.
VPS29	0.17	0.0000373	0.000846	12	Homo sapiens vacuolar protein sorting 29 homolog (S. cerevisiae) (VPS29), transcript variant 1, mRNA.
MRLC2	0.1	0.0000374	0.000846	18	Homo sapiens myosin regulatory light chain MRLC2 (MRLC2), mRNA.
FUCA1	0.14	0.0000374	0.000846	1	Homo sapiens fucosidase, alpha-L- 1, tissue (FUCA1), mRNA.
DPY30	0.18	0.0000379	0.000846	2	Homo sapiens dpy-30 homolog (C. elegans) (DPY30), mRNA.
SRP72	0.11	0.0000381	0.000846	4	Homo sapiens signal recognition particle 72kDa (SRP72), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
C14ORF138	0.17	0.0000383	0.000846	14	Homo sapiens chromosome 14 open reading frame 138 (C14orf138), transcript variant 1, mRNA.
UBE2N	0.14	0.0000383	0.000846	12	Homo sapiens ubiquitin-conjugating enzyme E2N (UBC13 homolog, yeast) (UBE2N), mRNA.
RSL24D1	0.42	0.0000388	0.000846	15	Homo sapiens ribosomal L24 domain containing 1 (RSL24D1), mRNA.
DSE	0.15	0.0000396	0.000853	6	Homo sapiens dermatan sulfate epimerase (DSE), transcript variant 1, mRNA.
RAB5A	0.1	0.0000397	0.000853	3	Homo sapiens RAB5A, member RAS oncogene family (RAB5A), mRNA.
RWDD1	0.21	0.0000404	0.000861	6	Homo sapiens RWD domain containing 1 (RWDD1), transcript variant 3, mRNA.
PPM1B	0.14	0.0000404	0.000861	2	Homo sapiens protein phosphatase 1B (formerly 2C), magnesium-dependent, beta isoform (PPM1B), transcript variant 2, mRNA.
TMED10P	0.1	0.0000412	0.000868	8	Homo sapiens transmembrane emp24-like trafficking protein 10 (yeast) pseudogene (TMED10P), non-coding RNA.
EFHA1	0.15	0.0000416	0.000868	13	Homo sapiens EF-hand domain family, member A1 (EFHA1), mRNA.
RHOT1	0.13	0.0000419	0.000868	17	Homo sapiens ras homolog gene family, member T1 (RHOT1), transcript variant 2, mRNA.
IRF2BP2	0.12	0.000042	0.000868	1	Homo sapiens interferon regulatory factor 2 binding protein 2 (IRF2BP2), transcript variant 1, mRNA.
ACADM	0.19	0.0000428	0.00088	1	Homo sapiens acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain (ACADM), nuclear gene encoding mitochondrial protein, mRNA.
KLHL5	0.1	0.0000446	0.000897	4	Homo sapiens kelch-like 5 (Drosophila) (KLHL5), transcript variant 3, mRNA.
SERP1	0.15	0.0000446	0.000897	3	Homo sapiens stress-associated endoplasmic reticulum protein 1 (SERP1), mRNA.
ATP5O	0.14	0.0000468	0.000929	21	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit (ATP5O), nuclear gene encoding mitochondrial protein, mRNA.
MDH1	0.1	0.0000474	0.000933	2	Homo sapiens malate dehydrogenase 1, NAD (soluble) (MDH1), mRNA.
PPP1R2	0.16	0.0000475	0.000933	3	Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 2 (PPP1R2), mRNA.
SS18L2	0.14	0.0000475	0.000933	3	Homo sapiens synovial sarcoma translocation gene on chromosome 18-like 2 (SS18L2), mRNA.
RPLP0	0.16	0.00005	0.000943	12	Homo sapiens ribosomal protein, large, P0 (RPLP0), transcript variant 1, mRNA.
NIF3L1	0.11	0.00005	0.000943	2	Homo sapiens NIF3 NGG1 interacting factor 3-like 1 (S. pombe) (NIF3L1), mRNA.
PSMB7	0.13	0.0000501	0.000943	9	Homo sapiens proteasome (prosome, macropain) subunit, beta type, 7 (PSMB7), mRNA.
PSMA3	0.17	0.0000502	0.000943	14	Homo sapiens proteasome (prosome, macropain) subunit, alpha type, 3 (PSMA3), transcript variant 2, mRNA.
ZFR	0.13	0.0000502	0.000943	5	Homo sapiens zinc finger RNA binding protein (ZFR), mRNA.
FOXN2	0.17	0.0000524	0.000964	2	Homo sapiens forkhead box N2 (FOXN2), mRNA.
FAR1	0.18	0.0000555	0.00101	11	Homo sapiens fatty acyl CoA reductase 1 (FAR1), mRNA.
EIF4E3	0.12	0.0000564	0.00101	3	Homo sapiens eukaryotic translation initiation factor 4E family member 3 (EIF4E3), mRNA.
RAB3GAP2	0.14	0.0000581	0.00103	1	Homo sapiens RAB3 GTPase activating protein subunit 2 (non-catalytic) (RAB3GAP2), mRNA.
KIAA1370	0.14	0.0000602	0.00104	15	Homo sapiens KIAA1370 (KIAA1370), mRNA.
SLMAP	0.13	0.0000602	0.00104	3	Homo sapiens sarcolemma associated protein (SLMAP), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
OXR1	0.24	0.0000613	0.00105	8	Homo sapiens oxidation resistance 1 (OXR1), mRNA.
RCOR3	0.17	0.0000615	0.00105	1	Homo sapiens REST corepressor 3 (RCOR3), mRNA.
FAM49B	0.1	0.0000643	0.00107	8	Homo sapiens family with sequence similarity 49, member B (FAM49B), mRNA.
RASA1	0.13	0.0000669	0.00111	5	Homo sapiens RAS p21 protein activator (GTPase activating protein) 1 (RASA1), transcript variant 1, mRNA.
ZNHIT3	0.23	0.0000678	0.00111	17	Homo sapiens zinc finger, HIT type 3 (ZNHIT3), mRNA.
C8ORF59	0.22	0.0000684	0.00111	8	Homo sapiens chromosome 8 open reading frame 59 (C8orf59), transcript variant 3, mRNA.
ATP5L	0.14	0.00007	0.00112	11	Homo sapiens ATP synthase, H+ transporting, mitochondrial F0 complex, subunit G (ATP5L), nuclear gene encoding mitochondrial protein, mRNA.
SP3	0.18	0.0000701	0.00112	2	Homo sapiens Sp3 transcription factor (SP3), transcript variant 2, mRNA.
GAPT	0.17	0.0000709	0.00113	5	Homo sapiens GRB2-binding adaptor protein, transmembrane (GAPT), mRNA.
CD302	0.16	0.0000715	0.00114	2	Homo sapiens CD302 molecule (CD302), mRNA.
SERPINB1	0.12	0.0000731	0.00114	6	Homo sapiens serpin peptidase inhibitor, clade B (ovalbumin), member 1 (SERPINB1), mRNA.
NIN	0.12	0.0000739	0.00115	14	Homo sapiens ninein (GSK3B interacting protein) (NIN), transcript variant 2, mRNA.
RPL7	0.36	0.0000757	0.00117	8	Homo sapiens ribosomal protein L7 (RPL7), mRNA.
NDUFA4	0.3	0.0000782	0.0012	7	Homo sapiens NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa (NDUFA4), nuclear gene encoding mitochondrial protein, mRNA.
PCNP	0.19	0.0000794	0.0012	3	Homo sapiens PEST proteolytic signal containing nuclear protein (PCNP), mRNA.
PSIP1	0.16	0.0000797	0.0012	9	Homo sapiens PC4 and SFRS1 interacting protein 1 (PSIP1), transcript variant 2, mRNA.
TMCO1	0.19	0.0000805	0.00121	1	Homo sapiens transmembrane and coiled-coil domains 1 (TMCO1), mRNA.
SRP14P1	0.15	0.0000822	0.00123	12	Homo sapiens signal recognition particle 14kDa (homologous Alu RNA binding protein) pseudogene 1 (SRP14P1), non-coding RNA.
RHEB	0.12	0.000083	0.00124	7	Homo sapiens Ras homolog enriched in brain (RHEB), mRNA.
SCAMP1	0.14	0.0000838	0.00124	5	Homo sapiens secretory carrier membrane protein 1 (SCAMP1), mRNA.
ATP5J	0.18	0.0000848	0.00125	21	Homo sapiens ATP synthase, H+ transporting, mitochondrial F0 complex, subunit F6 (ATP5J), nuclear gene encoding mitochondrial protein, transcript variant 3, mRNA.
UQCRH	0.14	0.0000858	0.00125	1	Homo sapiens ubiquinol-cytochrome c reductase hinge protein (UQCRH), mRNA.
RPL24	0.13	0.0000888	0.00128	3	Homo sapiens ribosomal protein L24 (RPL24), mRNA.
HINT1	0.34	0.0000901	0.00129	5	Homo sapiens histidine triad nucleotide binding protein 1 (HINT1), mRNA.
TMEM126B	0.21	0.0000901	0.00129	11	Homo sapiens transmembrane protein 126B (TMEM126B), mRNA.
TMED5	0.18	0.0000914	0.0013	1	Homo sapiens transmembrane emp24 protein transport domain containing 5 (TMED5), mRNA.
RPS9	0.13	0.0000918	0.0013	19	Homo sapiens ribosomal protein S9 (RPS9), mRNA.
STAG2	0.16	0.0000922	0.0013	X	Homo sapiens stromal antigen 2 (STAG2), transcript variant 2, mRNA.
PAPD4	0.1	0.000094	0.00131	5	Homo sapiens PAP associated domain containing 4 (PAPD4), mRNA.
ACSL4	0.17	0.0000969	0.00134	X	Homo sapiens acyl-CoA synthetase long-chain family member 4 (ACSL4), transcript variant 1, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
LYRM5	0.14	0.0000971	0.00134	12	Homo sapiens LYR motif containing 5 (LYRM5), mRNA.
GMFG	0.12	0.0000988	0.00135	19	Homo sapiens glia maturation factor, gamma (GMFG), mRNA.
ANKRD12	0.19	0.000102	0.00138	18	Homo sapiens ankyrin repeat domain 12 (ANKRD12), transcript variant 1, mRNA.
CAPZA2	0.21	0.000103	0.00139	7	Homo sapiens capping protein (actin filament) muscle Z-line, alpha 2 (CAPZA2), mRNA.
SLC40A1	0.13	0.000106	0.0014	2	Homo sapiens solute carrier family 40 (iron-regulated transporter), member 1 (SLC40A1), mRNA.
SRP14	0.14	0.000107	0.00141	15	Homo sapiens signal recognition particle 14kDa (homologous Alu RNA binding protein) (SRP14), mRNA.
FAM160B1	0.16	0.000107	0.00141	10	Homo sapiens family with sequence similarity 160, member B1 (FAM160B1), transcript variant 1, mRNA.
S100A9	0.1	0.000114	0.00146	1	Homo sapiens S100 calcium binding protein A9 (calgranulin B) (S100A9), mRNA.
FEM1C	0.15	0.000115	0.00147	5	Homo sapiens fem-1 homolog c (C. elegans) (FEM1C), mRNA.
KLF9	0.17	0.000118	0.0015	9	Homo sapiens Kruppel-like factor 9 (KLF9), mRNA.
GALC	0.1	0.000119	0.0015	14	Homo sapiens galactosylceramidase (GALC), transcript variant 1, mRNA.
COX7C	0.33	0.00012	0.0015	5	Homo sapiens cytochrome c oxidase subunit VIIc (COX7C), nuclear gene encoding mitochondrial protein, mRNA.
PNRC2	0.15	0.00012	0.0015	1	Homo sapiens proline-rich nuclear receptor coactivator 2 (PNRC2), mRNA.
RPS15A	0.34	0.000122	0.00151	16	Homo sapiens ribosomal protein S15a (RPS15A), transcript variant 1, mRNA.
ACN9	0.13	0.000124	0.00153	7	Homo sapiens ACN9 homolog (S. cerevisiae) (ACN9), mRNA.
ZRANB2	0.19	0.000124	0.00153	1	Homo sapiens zinc finger, RAN-binding domain containing 2 (ZRANB2), transcript variant 2, mRNA.
GLO1	0.11	0.000131	0.0016	6	Homo sapiens glyoxalase I (GLO1), mRNA.
PSMA6	0.23	0.000131	0.0016	14	Homo sapiens proteasome (prosome, macropain) subunit, alpha type, 6 (PSMA6), mRNA.
HMGN4	0.12	0.000132	0.0016	6	Homo sapiens high mobility group nucleosomal binding domain 4 (HMGN4), mRNA.
UBL5	0.15	0.000134	0.00161	19	Homo sapiens ubiquitin-like 5 (UBL5), transcript variant 2, mRNA.
CHD1	0.11	0.000136	0.00163	5	Homo sapiens chromodomain helicase DNA binding protein 1 (CHD1), mRNA.
TBCA	0.14	0.000137	0.00163	5	Homo sapiens tubulin folding cofactor A (TBCA), mRNA.
FBXO33	0.11	0.000137	0.00163	14	Homo sapiens F-box protein 33 (FBXO33), mRNA.
TRIM33	0.1	0.000141	0.00164	1	Homo sapiens tripartite motif-containing 33 (TRIM33), transcript variant a, mRNA.
PPP1CB	0.17	0.000141	0.00164	2	Homo sapiens protein phosphatase 1, catalytic subunit, beta isoform (PPP1CB), transcript variant 3, mRNA.
LSM1	0.15	0.000141	0.00164	8	Homo sapiens LSM1 homolog, U6 small nuclear RNA associated (S. cerevisiae) (LSM1), mRNA.
RPL15	0.11	0.000144	0.00166	3	Homo sapiens ribosomal protein L15 (RPL15), mRNA.
POLE4	0.12	0.000145	0.00167	2	Homo sapiens polymerase (DNA-directed), epsilon 4 (p12 subunit) (POLE4), mRNA.
MITD1	0.12	0.000148	0.00168	2	Homo sapiens MIT, microtubule interacting and transport, domain containing 1 (MITD1), mRNA.
STX7	0.13	0.000153	0.00173	6	Homo sapiens syntaxin 7 (STX7), mRNA.
NCK1	0.12	0.000153	0.00173	3	Homo sapiens NCK adaptor protein 1 (NCK1), mRNA.
DCP2	0.14	0.000154	0.00173	5	Homo sapiens DCP2 decapping enzyme homolog (S. cerevisiae) (DCP2), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
MGST3	0.12	0.000155	0.00173	1	Homo sapiens microsomal glutathione S-transferase 3 (MGST3), mRNA.
C2ORF64	0.12	0.000156	0.00173	2	Homo sapiens chromosome 2 open reading frame 64 (C2orf64), mRNA.
C20ORF52	0.13	0.000158	0.00175	20	Homo sapiens chromosome 20 open reading frame 52 (C20orf52), mRNA.
C17ORF61	0.13	0.000159	0.00175	17	Homo sapiens chromosome 17 open reading frame 61 (C17orf61), mRNA.
ZBED5	0.14	0.000159	0.00175	11	Homo sapiens zinc finger, BED-type containing 5 (ZBED5), mRNA.
MRPL11	0.12	0.000161	0.00177	11	Homo sapiens mitochondrial ribosomal protein L11 (MRPL11), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
PFAAP5	0.16	0.000162	0.00177	13	Homo sapiens phosphonoformate immuno-associated protein 5 (PFAAP5), mRNA.
PDCD10	0.27	0.000168	0.00182	3	Homo sapiens programmed cell death 10 (PDCD10), transcript variant 2, mRNA.
DDX59	0.11	0.00017	0.00183	1	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 59 (DDX59), transcript variant 1, mRNA.
CRBN	0.18	0.00017	0.00183	3	Homo sapiens cereblon (CRBN), mRNA.
MTDH	0.11	0.000174	0.00186	8	Homo sapiens metadherin (MTDH), mRNA.
DEK	0.16	0.000176	0.00186	6	Homo sapiens DEK oncogene (DNA binding) (DEK), mRNA.
RBX1	0.23	0.000177	0.00187	22	Homo sapiens ring-box 1 (RBX1), mRNA.
NAT5	0.1	0.000179	0.00188	20	Homo sapiens N-acetyltransferase 5 (GCN5-related, putative) (NAT5), transcript variant 3, mRNA.
DBI	0.23	0.000192	0.00195	2	Homo sapiens diazepam binding inhibitor (GABA receptor modulator, acyl-Coenzyme A binding protein) (DBI), transcript variant 2, mRNA.
RPS6KB1	0.15	0.000194	0.00196	17	Homo sapiens ribosomal protein S6 kinase, 70kDa, polypeptide 1 (RPS6KB1), mRNA.
ATP5EP2	0.1	0.000198	0.00199	13	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit pseudogene 2 (ATP5EP2), transcript variant 6, non-coding RNA.
NSL1	0.11	0.0002	0.002	1	Homo sapiens NSL1, MIND kinetochore complex component, homolog (S. cerevisiae) (NSL1), transcript variant 2, mRNA.
NME1-NME2	0.1	0.0002	0.002	17	Homo sapiens NME1-NME2 readthrough (NME1-NME2), mRNA.
PAIP2	0.1	0.000202	0.002	5	Homo sapiens poly(A) binding protein interacting protein 2 (PAIP2), transcript variant 1, mRNA.
EVI2A	0.34	0.000203	0.00201	17	Homo sapiens ecotropic viral integration site 2A (EVI2A), transcript variant 2, mRNA.
CHMP5	0.24	0.000206	0.00202	9	Homo sapiens chromatin modifying protein 5 (CHMP5), mRNA.
TMEM167B	0.1	0.000206	0.00202	1	Homo sapiens transmembrane protein 167B (TMEM167B), mRNA.
GCA	0.12	0.000208	0.00202	2	Homo sapiens grancalcin, EF-hand calcium binding protein (GCA), mRNA.
ZMPSTE24	0.1	0.000208	0.00202	1	Homo sapiens zinc metallopeptidase (STE24 homolog, S. cerevisiae) (ZMPSTE24), mRNA.
AP1S2	0.2	0.000209	0.00203	X	Homo sapiens adaptor-related protein complex 1, sigma 2 subunit (AP1S2), mRNA.
OBFC2A	0.12	0.000211	0.00203	2	Homo sapiens oligonucleotide/oligosaccharide-binding fold containing 2A (OBFC2A), mRNA.
EFR3A	0.14	0.000211	0.00203	8	Homo sapiens EFR3 homolog A (S. cerevisiae) (EFR3A), mRNA.
REEP5	0.13	0.000219	0.00208	5	Homo sapiens receptor accessory protein 5 (REEP5), mRNA.
RGS18	0.21	0.000221	0.0021	1	Homo sapiens regulator of G-protein signaling 18 (RGS18), mRNA.
CCDC53	0.12	0.000227	0.00214	12	Homo sapiens coiled-coil domain containing 53 (CCDC53), mRNA.
ADD3	0.11	0.000228	0.00214	10	Homo sapiens adducin 3 (gamma) (ADD3), transcript variant 3, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
COQ10B	0.1	0.000229	0.00214	2	Homo sapiens coenzyme Q10 homolog B (S. cerevisiae) (COQ10B), mRNA.
MRPS22	0.12	0.000229	0.00214	3	Homo sapiens mitochondrial ribosomal protein S22 (MRPS22), nuclear gene encoding mitochondrial protein, mRNA.
OSTC	0.14	0.000229	0.00214	4	Homo sapiens oligosaccharyltransferase complex subunit (OSTC), mRNA.
BTAF1	0.15	0.000232	0.00216	10	Homo sapiens BTAF1 RNA polymerase II, B-TFIID transcription factor-associated, 170kDa (Mot1 homolog, S. cerevisiae) (BTAF1), mRNA.
RPS3	0.1	0.000233	0.00216	11	Homo sapiens ribosomal protein S3 (RPS3), mRNA.
SNRPG	0.23	0.000237	0.00218	2	Homo sapiens small nuclear ribonucleoprotein polypeptide G (SNRPG), mRNA.
USP15	0.17	0.000247	0.00224	12	Homo sapiens ubiquitin specific peptidase 15 (USP15), mRNA.
MBNL1	0.16	0.000252	0.00228	3	Homo sapiens muscleblind-like (Drosophila) (MBNL1), transcript variant 6, mRNA.
PSMA4	0.2	0.000252	0.00228	15	Homo sapiens proteasome (prosome, macropain) subunit, alpha type, 4 (PSMA4), mRNA.
CBFB	0.12	0.000258	0.00232	16	Homo sapiens core-binding factor, beta subunit (CBFB), transcript variant 2, mRNA.
VPS36	0.15	0.000267	0.00237	13	Homo sapiens vacuolar protein sorting 36 homolog (S. cerevisiae) (VPS36), mRNA.
RPL17	0.34	0.000268	0.00237	18	Homo sapiens ribosomal protein L17 (RPL17), transcript variant 2, mRNA.
COX6C	0.21	0.000271	0.00237	8	Homo sapiens cytochrome c oxidase subunit VIc (COX6C), mRNA.
RPL9	0.37	0.000275	0.0024	4	Homo sapiens ribosomal protein L9 (RPL9), transcript variant 2, mRNA.
CMPK1	0.17	0.000283	0.00245	1	Homo sapiens cytidine monophosphate (UMP-CMP) kinase 1, cytosolic (CMPK1), mRNA.
ZSWIM6	0.11	0.000288	0.00247	5	PREDICTED: Homo sapiens zinc finger, SWIM-type containing 6 (ZSWIM6), mRNA.
RPL35	0.17	0.000288	0.00247	9	Homo sapiens ribosomal protein L35 (RPL35), mRNA.
ERH	0.16	0.000299	0.00252	14	Homo sapiens enhancer of rudimentary homolog (Drosophila) (ERH), mRNA.
FAM116A	0.12	0.000306	0.00258		PREDICTED: Homo sapiens family with sequence similarity 116, member A (FAM116A), mRNA.
ATP11B	0.11	0.000311	0.0026	3	Homo sapiens ATPase, class VI, type 11B (ATP11B), mRNA.
FAM60A	0.12	0.000316	0.00263	12	Homo sapiens family with sequence similarity 60, member A (FAM60A), transcript variant 1, mRNA.
SEC11C	0.18	0.000325	0.00269	18	Homo sapiens SEC11 homolog C (S. cerevisiae) (SEC11C), mRNA.
RSBN1	0.14	0.000331	0.00274	1	Homo sapiens round spermatid basic protein 1 (RSBN1), mRNA.
PTEN	0.1	0.000337	0.00276	10	Homo sapiens phosphatase and tensin homolog (PTEN), mRNA.
LACTB	0.1	0.000337	0.00276	15	Homo sapiens lactamase, beta (LACTB), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
PPA1	0.12	0.000341	0.00279	10	Homo sapiens pyrophosphatase (inorganic) 1 (PPA1), mRNA.
RPS27	0.22	0.000345	0.0028	1	Homo sapiens ribosomal protein S27 (metallopanstimulin 1) (RPS27), mRNA.
CGGBP1	0.14	0.000353	0.00282	3	Homo sapiens CGG triplet repeat binding protein 1 (CGGBP1), transcript variant 1, mRNA.
SEC11A	0.11	0.000353	0.00282	15	Homo sapiens SEC11 homolog A (S. cerevisiae) (SEC11A), mRNA.
RPS24	0.25	0.000359	0.00286	10	Homo sapiens ribosomal protein S24 (RPS24), transcript variant 1, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
C14ORF156	0.23	0.00037	0.00291	14	Homo sapiens chromosome 14 open reading frame 156 (C14orf156), mRNA.
TDG	0.1	0.00037	0.00291	12	Homo sapiens thymine-DNA glycosylase (TDG), mRNA.
GMFB	0.16	0.000385	0.00299	14	Homo sapiens glia maturation factor, beta (GMFB), mRNA.
NAE1	0.11	0.000386	0.00299	16	Homo sapiens NEDD8 activating enzyme E1 subunit 1 (NAE1), transcript variant 3, mRNA.
PRICKLE4	0.12	0.00039	0.003	6	Homo sapiens prickle homolog 4 (Drosophila) (PRICKLE4), mRNA.
RPL31	0.34	0.00039	0.003	2	Homo sapiens ribosomal protein L31 (RPL31), transcript variant 1, mRNA.
GPR65	0.17	0.000397	0.00303	14	Homo sapiens G protein-coupled receptor 65 (GPR65), mRNA.
IFRD1	0.15	0.00041	0.0031	7	Homo sapiens interferon-related developmental regulator 1 (IFRD1), transcript variant 1, mRNA.
C12ORF47	0.1	0.000415	0.00312		PREDICTED: Homo sapiens chromosome 12 open reading frame 47 (C12orf47), misc RNA.
CD48	0.13	0.00042	0.00312	1	Homo sapiens CD48 molecule (CD48), mRNA.
THOC7	0.1	0.000422	0.00312	3	Homo sapiens THO complex 7 homolog (Drosophila) (THOC7), mRNA.
CHMP1B	0.1	0.000424	0.00312	18	Homo sapiens chromatin modifying protein 1B (CHMP1B), mRNA.
TMEM123	0.18	0.000427	0.00312	11	Homo sapiens transmembrane protein 123 (TMEM123), mRNA.
SF3B1	0.1	0.00044	0.00319	2	Homo sapiens splicing factor 3b, subunit 1, 155kDa (SF3B1), transcript variant 1, mRNA.
TMEM14D	0.14	0.000444	0.00321	10	PREDICTED: Homo sapiens transmembrane protein 14D (TMEM14D), mRNA.
PRDM1	0.12	0.000456	0.00326	6	Homo sapiens PR domain containing 1, with ZNF domain (PRDM1), transcript variant 1, mRNA.
CTSH	0.1	0.00046	0.00327	15	Homo sapiens cathepsin H (CTSH), transcript variant 1, mRNA.
CD47	0.11	0.000505	0.00346	3	Homo sapiens CD47 molecule (CD47), transcript variant 2, mRNA.
UBE2J1	0.12	0.00051	0.00348	6	Homo sapiens ubiquitin-conjugating enzyme E2, J1 (UBC6 homolog, yeast) (UBE2J1), mRNA.
AKAP11	0.16	0.000518	0.00352	13	Homo sapiens A kinase (PRKA) anchor protein 11 (AKAP11), mRNA.
RPS18	0.18	0.00052	0.00352	6	Homo sapiens ribosomal protein S18 (RPS18), mRNA.
CASP1	0.12	0.000538	0.0036	11	Homo sapiens caspase 1, apoptosis-related cysteine peptidase (interleukin 1, beta, convertase) (CASP1), transcript variant delta, mRNA.
CCDC72	0.27	0.000547	0.00362	3	Homo sapiens coiled-coil domain containing 72 (CCDC72), mRNA.
UBE3A	0.11	0.000557	0.00365	15	Homo sapiens ubiquitin protein ligase E3A (UBE3A), transcript variant 2, mRNA.
CMTM6	0.1	0.000567	0.00369	3	Homo sapiens CKLF-like MARVEL transmembrane domain containing 6 (CMTM6), mRNA.
TAX1BP1	0.14	0.000568	0.00369	7	Homo sapiens Tax1 (human T-cell leukemia virus type I) binding protein 1 (TAX1BP1), transcript variant 2, mRNA.
SLC4A7	0.19	0.000571	0.0037	3	Homo sapiens solute carrier family 4, sodium bicarbonate cotransporter, member 7 (SLC4A7), mRNA.
UQCRHL	0.11	0.000572	0.0037	1	Homo sapiens ubiquinol-cytochrome c reductase hinge protein-like (UQCRHL), mRNA.
CLINT1	0.1	0.000582	0.00375	5	Homo sapiens clathrin interactor 1 (CLINT1), mRNA.
KRCC1	0.14	0.000586	0.00377	2	Homo sapiens lysine-rich coiled-coil 1 (KRCC1), mRNA.
RPL26	0.31	0.000587	0.00377	17	Homo sapiens ribosomal protein L26 (RPL26), mRNA.
ROMO1	0.13	0.000591	0.00377	20	Homo sapiens reactive oxygen species modulator 1 (ROMO1), nuclear gene encoding mitochondrial protein, mRNA.
PCMTD1	0.17	0.000595	0.00378	8	Homo sapiens protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1 (PCMTD1), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
ATP5I	0.1	0.000603	0.00382	4	Homo sapiens ATP synthase, H+ transporting, mitochondrial F0 complex, subunit E (ATP5I), nuclear gene encoding mitochondrial protein, mRNA.
WSB2	0.12	0.000607	0.00382	12	Homo sapiens WD repeat and SOCS box-containing 2 (WSB2), mRNA.
SNRPD2	0.13	0.000614	0.00385	19	Homo sapiens small nuclear ribonucleoprotein D2 polypeptide 16.5kDa (SNRPD2), transcript variant 1, mRNA.
USP16	0.11	0.000615	0.00385	21	Homo sapiens ubiquitin specific peptidase 16 (USP16), transcript variant 1, mRNA.
MTMR11	0.14	0.000619	0.00387	1	Homo sapiens myotubularin related protein 11 (MTMR11), mRNA.
RPL39	0.24	0.000621	0.00388	X	Homo sapiens ribosomal protein L39 (RPL39), mRNA.
C7ORF23	0.15	0.000625	0.00388	7	Homo sapiens chromosome 7 open reading frame 23 (C7orf23), mRNA.
HPRT1	0.1	0.000632	0.00392	X	Homo sapiens hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome) (HPRT1), mRNA.
MTIF3	0.11	0.000641	0.00395	13	Homo sapiens mitochondrial translational initiation factor 3 (MTIF3), nuclear gene encoding mitochondrial protein, mRNA.
BAZ2B	0.14	0.000659	0.00403	2	Homo sapiens bromodomain adjacent to zinc finger domain, 2B (BAZ2B), mRNA.
NDUFS5	0.14	0.00066	0.00403	1	Homo sapiens NADH dehydrogenase (ubiquinone) Fe-S protein 5, 15kDa (NADH-coenzyme Q reductase) (NDUFS5), mRNA.
TOMM7	0.19	0.000662	0.00403	7	Homo sapiens translocase of outer mitochondrial membrane 7 homolog (yeast) (TOMM7), nuclear gene encoding mitochondrial protein, mRNA.
C18ORF25	0.11	0.000665	0.00405	18	Homo sapiens chromosome 18 open reading frame 25 (C18orf25), transcript variant 2, mRNA.
TMEM50B	0.13	0.000675	0.00409	21	Homo sapiens transmembrane protein 50B (TMEM50B), mRNA.
RPS17	0.23	0.000694	0.00418	15	Homo sapiens ribosomal protein S17 (RPS17), mRNA.
ITGAE	0.1	0.000708	0.00424	17	Homo sapiens integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide) (ITGAE), mRNA.
ATP6V1G1	0.14	0.000709	0.00424	9	Homo sapiens ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G1 (ATP6V1G1), mRNA.
C14ORF106	0.12	0.000711	0.00424	14	Homo sapiens chromosome 14 open reading frame 106 (C14orf106), mRNA.
SMEK2	0.11	0.000716	0.00427	2	Homo sapiens SMEK homolog 2, suppressor of mek1 (Dictyostelium) (SMEK2), mRNA.
RPL23	0.31	0.000717	0.00427	17	Homo sapiens ribosomal protein L23 (RPL23), mRNA.
RPS6P1	0.17	0.000726	0.00431		PREDICTED: Homo sapiens misc_RNA (RPS6P1), miscRNA.
RPL13A	0.13	0.000737	0.00435	19	Homo sapiens ribosomal protein L13a (RPL13A), mRNA.
MRPS21	0.14	0.000747	0.00438	1	Homo sapiens mitochondrial ribosomal protein S21 (MRPS21), nuclear gene encoding mitochondrial protein, transcript variant 2, mRNA.
BIRC2	0.14	0.000756	0.00442	11	Homo sapiens baculoviral IAP repeat-containing 2 (BIRC2), mRNA.
SUZ12	0.12	0.000783	0.00451	17	Homo sapiens suppressor of zeste 12 homolog (Drosophila) (SUZ12), mRNA.
RPL5	0.11	0.000799	0.00456	1	Homo sapiens ribosomal protein L5 (RPL5), mRNA.
TMED2	0.1	0.000806	0.0046	12	Homo sapiens transmembrane emp24 domain trafficking protein 2 (TMED2), mRNA.
RPS10	0.13	0.000808	0.0046	6	Homo sapiens ribosomal protein S10 (RPS10), mRNA.
CENTB2	0.1	0.000811	0.0046	3	Homo sapiens centaurin, beta 2 (CENTB2), mRNA.
SNX14	0.12	0.000816	0.00461	6	Homo sapiens sorting nexin 14 (SNX14), transcript variant 2, mRNA.
RAB8B	0.12	0.000829	0.00467	15	Homo sapiens RAB8B, member RAS oncogene family (RAB8B), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
SFRS2IP	0.1	0.000834	0.00467	12	Homo sapiens splicing factor, arginine/serine-rich 2, interacting protein (SFRS2IP), mRNA.
CUL4A	0.12	0.000834	0.00467	13	Homo sapiens cullin 4A (CUL4A), transcript variant 2, mRNA.
COQ5	0.11	0.00084	0.00469	12	Homo sapiens coenzyme Q5 homolog, methyltransferase (S. cerevisiae) (COQ5), mRNA.
MS4A6A	0.14	0.000845	0.00471	11	Homo sapiens membrane-spanning 4-domains, subfamily A, member 6A (MS4A6A), transcript variant 3, mRNA.
NDUFA1	0.1	0.000849	0.00472	X	Homo sapiens NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa (NDUFA1), nuclear gene encoding mitochondrial protein, mRNA.
MATR3	0.13	0.000851	0.00473	5	Homo sapiens matrin 3 (MATR3), transcript variant 1, mRNA.
RPS3A	0.31	0.000861	0.00475	4	Homo sapiens ribosomal protein S3A (RPS3A), mRNA.
PQLC3	0.11	0.000883	0.00483	2	Homo sapiens PQ loop repeat containing 3 (PQLC3), mRNA.
NDUFB3	0.15	0.000888	0.00483	2	Homo sapiens NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 3, 12kDa (NDUFB3), mRNA.
SAMD9	0.14	0.000896	0.00484	7	Homo sapiens sterile alpha motif domain containing 9 (SAMD9), mRNA.
NDUFB5	0.11	0.000899	0.00484	3	Homo sapiens NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa (NDUFB5), nuclear gene encoding mitochondrial protein, mRNA.
RPS6	0.13	0.000901	0.00484	9	Homo sapiens ribosomal protein S6 (RPS6), mRNA.
GNAI3	0.12	0.000924	0.0049	1	Homo sapiens guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3 (GNAI3), mRNA.
NDUFB7	0.12	0.000927	0.0049	19	Homo sapiens NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 7, 18kDa (NDUFB7), nuclear gene encoding mitochondrial protein, mRNA.
CXCR4	0.1	0.000928	0.0049	2	Homo sapiens chemokine (C-X-C motif) receptor 4 (CXCR4), transcript variant 1, mRNA.
LEMD3	0.15	0.000934	0.00491	12	Homo sapiens LEM domain containing 3 (LEMD3), mRNA.
TRAM1	0.11	0.000938	0.00492	8	Homo sapiens translocation associated membrane protein 1 (TRAM1), mRNA.
HIF1A	0.11	0.00096	0.00499	14	Homo sapiens hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor) (HIF1A), transcript variant 2, mRNA.
C1ORF59	0.1	0.000989	0.00507	1	Homo sapiens chromosome 1 open reading frame 59 (C1orf59), mRNA.
RPA3	0.11	0.00101	0.00517	7	Homo sapiens replication protein A3, 14kDa (RPA3), mRNA.
WDFY1	0.12	0.00104	0.00524	2	Homo sapiens WD repeat and FYVE domain containing 1 (WDFY1), mRNA.
VPS26A	0.11	0.00104	0.00525	10	Homo sapiens vacuolar protein sorting 26 homolog A (S. pombe) (VPS26A), transcript variant 2, mRNA.
RPL35A	0.15	0.00106	0.00533	3	Homo sapiens ribosomal protein L35a (RPL35A), mRNA.
RPL13	0.13	0.00109	0.00539	16	Homo sapiens ribosomal protein L13 (RPL13), transcript variant 2, mRNA.
SNHG5	0.27	0.0011	0.00542	6	Homo sapiens small nucleolar RNA host gene (non-protein coding) 5 (SNHG5) on chromosome 6.
MRFAP1L1	0.1	0.0011	0.00542	4	Homo sapiens Morf4 family associated protein 1-like 1 (MRFAP1L1), transcript variant 2, mRNA.
CLK1	0.15	0.00111	0.00544	2	Homo sapiens CDC-like kinase 1 (CLK1), mRNA.
CREG1	0.13	0.00113	0.0055	1	Homo sapiens cellular repressor of E1A-stimulated genes 1 (CREG1), mRNA.
RDH14	0.12	0.00117	0.00559	2	Homo sapiens retinol dehydrogenase 14 (all-trans/9-cis/11-cis) (RDH14), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
PELI1	0.11	0.00121	0.00575	2	Homo sapiens pellino homolog 1 (Drosophila) (PELI1), mRNA.
RAP2C	0.12	0.00122	0.00577	X	Homo sapiens RAP2C, member of RAS oncogene family (RAP2C), mRNA.
TNFAIP8	0.12	0.00123	0.00579	5	Homo sapiens tumor necrosis factor, alpha-induced protein 8 (TNFAIP8), transcript variant 2, mRNA.
FAU	0.13	0.00123	0.00579	11	Homo sapiens Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (FAU), mRNA.
SLC38A2	0.14	0.00127	0.00594	12	Homo sapiens solute carrier family 38, member 2 (SLC38A2), mRNA.
COPS8	0.1	0.00128	0.00595	2	Homo sapiens COP9 constitutive photomorphogenic homolog subunit 8 (Arabidopsis) (COPS8), transcript variant 2, mRNA.
TMEM188	0.1	0.00129	0.00599	16	Homo sapiens transmembrane protein 188 (TMEM188), mRNA.
DYNLT1	0.1	0.00134	0.00616	6	Homo sapiens dynein, light chain, Tctex-type 1 (DYNLT1), mRNA.
FLJ31306	0.11	0.00137	0.00627	14	PREDICTED: Homo sapiens hypothetical protein FLJ31306 (FLJ31306), mRNA.
NSA2	0.12	0.0014	0.00639	5	Homo sapiens NSA2 ribosome biogenesis homolog (S. cerevisiae) (NSA2), mRNA.
PARK7	0.13	0.00145	0.00651	1	Homo sapiens Parkinson disease (autosomal recessive, early onset) 7 (PARK7), mRNA.
RPL27	0.19	0.00149	0.00665	17	Homo sapiens ribosomal protein L27 (RPL27), mRNA.
HMGB1L1	0.16	0.0015	0.00667	20	Homo sapiens high-mobility group box 1-like 1 (HMGB1L1), mRNA.
UBE2Q2	0.13	0.00156	0.00685	15	Homo sapiens ubiquitin-conjugating enzyme E2Q family member 2 (UBE2Q2), mRNA.
RPS29	0.12	0.00158	0.00692	14	Homo sapiens ribosomal protein S29 (RPS29), transcript variant 1, mRNA.
CLEC12A	0.29	0.00159	0.00695	12	Homo sapiens C-type lectin domain family 12, member A (CLEC12A), transcript variant 2, mRNA.
THUMPD1	0.12	0.0016	0.00697	16	Homo sapiens THUMP domain containing 1 (THUMPD1), mRNA.
C16ORF61	0.14	0.0016	0.00698	16	Homo sapiens chromosome 16 open reading frame 61 (C16orf61), mRNA.
EEF1B2	0.19	0.00162	0.00702	2	Homo sapiens eukaryotic translation elongation factor 1 beta 2 (EEF1B2), transcript variant 3, mRNA.
NOP58	0.12	0.00165	0.0071	2	Homo sapiens NOP58 ribonucleoprotein homolog (yeast) (NOP58), mRNA.
SNX10	0.12	0.00167	0.00718	7	Homo sapiens sorting nexin 10 (SNX10), mRNA.
RPL30	0.13	0.00171	0.0073	8	Homo sapiens ribosomal protein L30 (RPL30), mRNA.
NME1	0.11	0.00176	0.00747	17	Homo sapiens non-metastatic cells 1, protein (NM23A) expressed in (NME1), transcript variant 2, mRNA.
PPIL3	0.14	0.00178	0.00753	2	Homo sapiens peptidylprolyl isomerase (cyclophilin)-like 3 (PPIL3), transcript variant PPIL3b, mRNA.
LSM5	0.12	0.00182	0.00765	7	Homo sapiens LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae) (LSM5), mRNA.
AIF1	0.1	0.00185	0.00773	6	Homo sapiens allograft inflammatory factor 1 (AIF1), transcript variant 1, mRNA.
RNASE2	0.27	0.00187	0.00777	14	Homo sapiens ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin) (RNASE2), mRNA.
DPYD	0.1	0.00191	0.00787	1	Homo sapiens dihydropyrimidine dehydrogenase (DPYD), transcript variant 1, mRNA.
ACAP2	0.1	0.00192	0.00789	3	Homo sapiens ArfGAP with coiled-coil, ankyrin repeat and PH domains 2 (ACAP2), mRNA.
BNIP3L	0.2	0.00194	0.00795	8	Homo sapiens BCL2/adenovirus E1B 19kDa interacting protein 3-like (BNIP3L), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
NDUFB2	0.13	0.00195	0.00797	7	Homo sapiens NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2, 8kDa (NDUFB2), nuclear gene encoding mitochondrial protein, mRNA.
RFX7	0.13	0.00197	0.00801	15	Homo sapiens regulatory factor X, 7 (RFX7), mRNA.
CD36	0.21	0.00201	0.00812	7	Homo sapiens CD36 molecule (thrombospondin receptor) (CD36), transcript variant 3, mRNA.
CD46	0.11	0.00217	0.00859	1	Homo sapiens CD46 molecule, complement regulatory protein (CD46), transcript variant m, mRNA.
CCDC14	0.11	0.0022	0.00867	3	Homo sapiens coiled-coil domain containing 14 (CCDC14), mRNA.
EEF1AL7	0.11	0.00222	0.00874	4	Homo sapiens eukaryotic translation elongation factor 1 alpha-like 7 (EEF1AL7), non-coding RNA.
ACP1	0.12	0.00223	0.00874	2	Homo sapiens acid phosphatase 1, soluble (ACP1), transcript variant 4, mRNA.
PRDX4	0.11	0.00229	0.00888	X	Homo sapiens peroxiredoxin 4 (PRDX4), mRNA.
RPS14	0.11	0.00232	0.00897	5	Homo sapiens ribosomal protein S14 (RPS14), transcript variant 2, mRNA.
ARGLU1	0.13	0.00234	0.00901	13	Homo sapiens arginine and glutamate rich 1 (ARGLU1), mRNA.
RBM25	0.1	0.00239	0.00917	14	Homo sapiens RNA binding motif protein 25 (RBM25), mRNA.
C16ORF63	0.12	0.00244	0.00931	16	Homo sapiens chromosome 16 open reading frame 63 (C16orf63), mRNA.
ITM2A	0.13	0.00244	0.00931	X	Homo sapiens integral membrane protein 2A (ITM2A), mRNA.
ASGR1	0.12	0.00244	0.00932	17	Homo sapiens asialoglycoprotein receptor 1 (ASGR1), mRNA.
CD52	0.19	0.00246	0.00932	1	Homo sapiens CD52 molecule (CD52), mRNA.
BMI1	0.14	0.00261	0.00973	10	Homo sapiens BMI1 polycomb ring finger oncogene (BMI1), mRNA.
PSTPIP2	0.11	0.00263	0.00979	18	Homo sapiens proline-serine-threonine phosphatase interacting protein 2 (PSTPIP2), mRNA.
TINP1	0.12	0.00265	0.00983	5	Homo sapiens TGF beta-inducible nuclear protein 1 (TINP1), mRNA.
KAT2B	0.13	0.00269	0.00991	3	Homo sapiens K(lysine) acetyltransferase 2B (KAT2B), mRNA.
SNORA25	0.1	0.00282	0.0103	11	Homo sapiens small nucleolar RNA, H/ACA box 25 (SNORA25), small nucleolar RNA.
SNRPF	0.1	0.00283	0.0103	12	Homo sapiens small nuclear ribonucleoprotein polypeptide F (SNRPF), mRNA.
RPL41	0.14	0.00283	0.0103	12	Homo sapiens ribosomal protein L41 (RPL41), transcript variant 2, mRNA.
JAZF1	0.11	0.00284	0.0103	7	Homo sapiens JAZF zinc finger 1 (JAZF1), mRNA.
NSMCE4A	0.1	0.00286	0.0104	10	Homo sapiens non-SMC element 4 homolog A (S. cerevisiae) (NSMCE4A), mRNA.
RPL11	0.12	0.00292	0.0106	1	Homo sapiens ribosomal protein L11 (RPL11), mRNA.
CROP	0.12	0.00295	0.0106	17	Homo sapiens cisplatin resistance-associated overexpressed protein (CROP), transcript variant 2, mRNA.
MST4	0.1	0.00303	0.0109	X	Homo sapiens serine/threonine protein kinase MST4 (MST4), transcript variant 1, mRNA.
ATP5E	0.14	0.0031	0.0111	20	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, epsilon subunit (ATP5E), nuclear gene encoding mitochondrial protein, mRNA.
CYBB	0.11	0.00317	0.0113	X	Homo sapiens cytochrome b-245, beta polypeptide (chronic granulomatous disease) (CYBB), mRNA.
PRKRIR	0.13	0.00322	0.0113	11	Homo sapiens protein-kinase, interferon-inducible double stranded RNA dependent inhibitor, repressor of (P58 repressor) (PRKRIR), mRNA.
HEBP2	0.12	0.00326	0.0115	6	Homo sapiens heme binding protein 2 (HEBP2), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
PTGER2	0.11	0.0033	0.0115	14	Homo sapiens prostaglandin E receptor 2 (subtype EP2), 53kDa (PTGER2), mRNA.
CD3D	0.13	0.00333	0.0116	11	Homo sapiens CD3d molecule, delta (CD3-TCR complex) (CD3D), transcript variant 2, mRNA.
TOMM6	0.1	0.00335	0.0117	6	Homo sapiens translocase of outer mitochondrial membrane 6 homolog (yeast) (TOMM6), nuclear gene encoding mitochondrial protein, mRNA.
SSB	0.1	0.00352	0.0121	2	Homo sapiens Sjogren syndrome antigen B (autoantigen La) (SSB), mRNA.
CNIH	0.12	0.00365	0.0125	14	Homo sapiens cornichon homolog (Drosophila) (CNIH), transcript variant 2, mRNA.
TMEM14C	0.14	0.00366	0.0125	6	Homo sapiens transmembrane protein 14C (TMEM14C), mRNA.
FYTTD1	0.1	0.00388	0.0131	3	Homo sapiens forty-two-three domain containing 1 (FYTTD1), transcript variant 2, mRNA.
SELK	0.12	0.00394	0.0132	3	Homo sapiens selenoprotein K (SELK), mRNA.
TMEM181	0.1	0.00411	0.0136	6	Homo sapiens transmembrane protein 181 (TMEM181), mRNA.
STAMBPL1	0.1	0.00416	0.0137	10	Homo sapiens STAM binding protein-like 1 (STAMBPL1), mRNA.
ZBTB33	0.11	0.00416	0.0137	X	Homo sapiens zinc finger and BTB domain containing 33 (ZBTB33), mRNA.
PFDN5	0.18	0.00465	0.015	12	Homo sapiens prefoldin subunit 5 (PFDN5), transcript variant 3, mRNA.
MGC87895	0.11	0.00474	0.0151		PREDICTED: Homo sapiens similar to ribosomal protein S14 (MGC87895), mRNA.
C20ORF199	0.13	0.00496	0.0156	20	Homo sapiens chromosome 20 open reading frame 199 (C20orf199), transcript variant 3, non-coding RNA.
RPL4	0.1	0.00505	0.0158	15	Homo sapiens ribosomal protein L4 (RPL4), mRNA.
DNAJB14	0.12	0.00522	0.0162	4	Homo sapiens DnaJ (Hsp40) homolog, subfamily B, member 14 (DNAJB14), transcript variant 2, mRNA.
PTPN12	0.1	0.00598	0.018	7	Homo sapiens protein tyrosine phosphatase, non-receptor type 12 (PTPN12), mRNA.
SACM1L	0.11	0.00621	0.0185	3	Homo sapiens SAC1 suppressor of actin mutations 1-like (yeast) (SACM1L), mRNA.
PGLYRP1	0.2	0.00627	0.0186	19	Homo sapiens peptidoglycan recognition protein 1 (PGLYRP1), mRNA.
CPD	0.1	0.00688	0.0199	17	Homo sapiens carboxypeptidase D (CPD), mRNA.
HIST1H4C	0.14	0.00699	0.0201	6	Homo sapiens histone cluster 1, H4c (HIST1H4C), mRNA.
C6ORF160	0.15	0.00772	0.0218		PREDICTED: Homo sapiens chromosome 6 open reading frame 160, transcript variant 4 (C6orf160), mRNA.
MRPS18C	0.12	0.00819	0.0228	4	Homo sapiens mitochondrial ribosomal protein S18C (MRPS18C), nuclear gene encoding mitochondrial protein, mRNA.
RAP1BL	0.11	0.00921	0.025		Homo sapiens hCG1757335 (RAP1BL), mRNA.
C17ORF45	0.13	0.00929	0.0252	17	Homo sapiens chromosome 17 open reading frame 45 (C17orf45), mRNA.
RPL21	0.15	0.00985	0.0265	13	Homo sapiens ribosomal protein L21 (RPL21), mRNA.
GIMAP2	0.19	0.01	0.0268	7	Homo sapiens GTPase, IMAP family member 2 (GIMAP2), mRNA.
KCNJ2	0.11	0.0101	0.027	17	Homo sapiens potassium inwardly-rectifying channel, subfamily J, member 2 (KCNJ2), mRNA.
SH2D1A	0.11	0.0104	0.0275	X	Homo sapiens SH2 domain protein 1A, Duncan's disease (lymphoproliferative syndrome) (SH2D1A), mRNA.
TNFAIP6	0.17	0.0117	0.0302	2	Homo sapiens tumor necrosis factor, alpha-induced protein 6 (TNFAIP6), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
C12ORF57	0.11	0.012	0.0309	12	Homo sapiens chromosome 12 open reading frame 57 (C12orf57), mRNA.
CD93	0.1	0.0124	0.0315	20	Homo sapiens CD93 molecule (CD93), mRNA.
CREB5	0.12	0.0132	0.033	7	Homo sapiens cAMP responsive element binding protein 5 (CREB5), transcript variant 1, mRNA.
PTGS2	0.16	0.014	0.0345	1	Homo sapiens prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase) (PTGS2), mRNA.
ANXA1	0.11	0.0145	0.0353	9	Homo sapiens annexin A1 (ANXA1), mRNA.
PLEKHA1	0.1	0.0156	0.0375	10	Homo sapiens pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 1 (PLEKHA1), transcript variant 2, mRNA.
C5ORF32	0.1	0.0166	0.0393	5	Homo sapiens chromosome 5 open reading frame 32 (C5orf32), mRNA.
IL1B	0.1	0.0192	0.0441	2	Homo sapiens interleukin 1, beta (IL1B), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
Down-regulated					
HNRNPUL2	-0.18	0.000000226	0.000206	11	Homo sapiens heterogeneous nuclear ribonucleoprotein U-like 2 (HNRNPUL2), mRNA.
RBM14	-0.12	0.000000251	0.000206	11	Homo sapiens RNA binding motif protein 14 (RBM14), mRNA.
TMEM69	-0.11	0.00000086	0.000276	1	Homo sapiens transmembrane protein 69 (TMEM69), mRNA.
SCAP	-0.13	0.00000249	0.000367	3	Homo sapiens SREBF chaperone (SCAP), mRNA.
FAM110A	-0.15	0.0000025	0.000367	20	Homo sapiens family with sequence similarity 110, member A (FAM110A), transcript variant 3, mRNA.
RBM10	-0.11	0.00000263	0.000367	X	Homo sapiens RNA binding motif protein 10 (RBM10), transcript variant 2, mRNA.
ZNF296	-0.15	0.00000363	0.000399	19	Homo sapiens zinc finger protein 296 (ZNF296), mRNA.
PNPT1	-0.11	0.00000424	0.000433	2	Homo sapiens polyribonucleotide nucleotidyltransferase 1 (PNPT1), mRNA.
RASGRP2	-0.11	0.00000439	0.000433	11	Homo sapiens RAS guanyl releasing protein 2 (calcium and DAG-regulated) (RASGRP2), transcript variant 1, mRNA.
DENND4B	-0.11	0.00000505	0.000436	1	Homo sapiens DENN/MADD domain containing 4B (DENND4B), mRNA.
ZC3H5	-0.1	0.00000614	0.000466		PREDICTED: Homo sapiens zinc finger CCCH-type containing 5 (ZC3H5), mRNA.
CXXC1	-0.11	0.00000662	0.000466	18	Homo sapiens CXXC finger 1 (PHD domain) (CXXC1), mRNA.
SUPT5H	-0.13	0.00000717	0.000485	19	Homo sapiens suppressor of Ty 5 homolog (S. cerevisiae) (SUPT5H), mRNA.
GANAB	-0.11	0.00000814	0.00052	11	Homo sapiens glucosidase, alpha; neutral AB (GANAB), transcript variant 2, mRNA.
PHF15	-0.13	0.00000866	0.000529	5	Homo sapiens PHD finger protein 15 (PHF15), mRNA.
KIAA1267	-0.1	0.00000931	0.000533	17	Homo sapiens KIAA1267 (KIAA1267), mRNA.
CLSTN1	-0.15	0.00000958	0.000533	1	Homo sapiens calsyntenin 1 (CLSTN1), transcript variant 1, mRNA.
POM121C	-0.12	0.00001	0.000533	7	Homo sapiens POM121 membrane glycoprotein C (POM121C), mRNA.
TSSC4	-0.1	0.00001	0.000533	11	Homo sapiens tumor suppressing subtransferable candidate 4 (TSSC4), mRNA.
UBQLN4	-0.11	0.0000104	0.000533	1	Homo sapiens ubiquilin 4 (UBQLN4), mRNA.
RANGAP1	-0.12	0.0000105	0.000533	22	Homo sapiens Ran GTPase activating protein 1 (RANGAP1), mRNA.
OSBPL7	-0.13	0.0000106	0.000533	17	Homo sapiens oxysterol binding protein-like 7 (OSBPL7), transcript variant 1, mRNA.
WDR23	-0.11	0.0000118	0.000558	14	Homo sapiens WD repeat domain 23 (WDR23), transcript variant 1, mRNA.
FBXO46	-0.16	0.0000122	0.000563	19	PREDICTED: Homo sapiens F-box protein 46, transcript variant 5 (FBXO46), mRNA.
PRKD2	-0.14	0.0000123	0.000563	19	Homo sapiens protein kinase D2 (PRKD2), mRNA.
VAMP2	-0.11	0.0000128	0.000575	17	Homo sapiens vesicle-associated membrane protein 2 (synaptobrevin 2) (VAMP2), mRNA.
NUMA1	-0.12	0.0000132	0.000579	11	Homo sapiens nuclear mitotic apparatus protein 1 (NUMA1), mRNA.
ST3GAL1	-0.13	0.0000136	0.000587	8	Homo sapiens ST3 beta-galactoside alpha-2,3-sialyltransferase 1 (ST3GAL1), transcript variant 1, mRNA.
EDC4	-0.1	0.0000138	0.000587	16	Homo sapiens enhancer of mRNA decapping 4 (EDC4), mRNA.
SIK3	-0.11	0.000016	0.000653	11	Homo sapiens SIK family kinase 3 (SIK3), mRNA.
CD97	-0.16	0.0000166	0.000658	19	Homo sapiens CD97 molecule (CD97), transcript variant 1, mRNA.
TLN1	-0.2	0.0000174	0.000676	9	Homo sapiens talin 1 (TLN1), mRNA.
STIP1	-0.13	0.0000177	0.000682	11	Homo sapiens stress-induced-phosphoprotein 1 (Hsp70/Hsp90-organizing protein) (STIP1), mRNA.
ITPKB	-0.12	0.0000182	0.000693	1	Homo sapiens inositol 1,4,5-trisphosphate 3-kinase B (ITPKB), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
RAB35	-0.1	0.0000189	0.000696	12	Homo sapiens RAB35, member RAS oncogene family (RAB35), mRNA.
RAB11FIP1	-0.13	0.0000198	0.000696	8	Homo sapiens RAB11 family interacting protein 1 (class I) (RAB11FIP1), transcript variant 3, mRNA.
RASAL3	-0.11	0.00002	0.000696	19	Homo sapiens RAS protein activator like 3 (RASAL3), mRNA.
WASF2	-0.17	0.0000206	0.000703	1	Homo sapiens WAS protein family, member 2 (WASF2), mRNA.
PHRF1	-0.11	0.0000207	0.000703	11	Homo sapiens PHD and ring finger domains 1 (PHRF1), mRNA.
ICAM2	-0.1	0.0000229	0.000726	17	Homo sapiens intercellular adhesion molecule 2 (ICAM2), transcript variant 1, mRNA.
HGS	-0.1	0.0000233	0.000726	17	Homo sapiens hepatocyte growth factor-regulated tyrosine kinase substrate (HGS), mRNA.
MEF2D	-0.11	0.000024	0.000726	1	Homo sapiens myocyte enhancer factor 2D (MEF2D), mRNA.
SPG7	-0.1	0.0000277	0.000759	16	Homo sapiens spastic paraplegia 7 (pure and complicated autosomal recessive) (SPG7), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA.
XRCC6	-0.12	0.000029	0.000771	22	Homo sapiens X-ray repair complementing defective repair in Chinese hamster cells 6 (XRCC6), mRNA.
FYN	-0.1	0.0000297	0.000777	6	Homo sapiens FYN oncogene related to SRC, FGR, YES (FYN), transcript variant 2, mRNA.
PDPR	-0.2	0.0000303	0.000784		PREDICTED: Homo sapiens pyruvate dehydrogenase phosphatase regulatory subunit (PDPR), mRNA.
TRIM28	-0.13	0.0000319	0.000807	19	Homo sapiens tripartite motif-containing 28 (TRIM28), mRNA.
AKAP13	-0.11	0.0000345	0.00084	15	Homo sapiens A kinase (PRKA) anchor protein 13 (AKAP13), transcript variant 2, mRNA.
AP1G2	-0.1	0.0000348	0.000844	14	Homo sapiens adaptor-related protein complex 1, gamma 2 subunit (AP1G2), mRNA.
MED25	-0.2	0.000035	0.000845	19	Homo sapiens mediator complex subunit 25 (MED25), mRNA.
GP9	-0.31	0.0000369	0.000846	3	Homo sapiens glycoprotein IX (platelet) (GP9), mRNA.
PRF1	-0.25	0.0000375	0.000846	10	Homo sapiens perforin 1 (pore forming protein) (PRF1), transcript variant 1, mRNA.
LBA1	-0.17	0.0000375	0.000846	3	Homo sapiens lupus brain antigen 1 (LBA1), mRNA.
PBX2	-0.12	0.0000378	0.000846	6	Homo sapiens pre-B-cell leukemia homeobox 2 (PBX2), mRNA.
YY1AP1	-0.14	0.0000383	0.000846	1	Homo sapiens YY1 associated protein 1 (YY1AP1), transcript variant 2, mRNA.
CLCN7	-0.12	0.0000386	0.000846	16	Homo sapiens chloride channel 7 (CLCN7), mRNA.
UBA52	-0.1	0.0000409	0.000868	19	Homo sapiens ubiquitin A-52 residue ribosomal protein fusion product 1 (UBA52), transcript variant 1, mRNA.
CDK5RAP3	-0.1	0.000042	0.000868	17	Homo sapiens CDK5 regulatory subunit associated protein 3 (CDK5RAP3), mRNA.
PACS1	-0.16	0.0000431	0.000882	11	Homo sapiens phosphofurin acidic cluster sorting protein 1 (PACS1), mRNA.
C20ORF55	-0.15	0.0000433	0.000882	20	Homo sapiens chromosome 20 open reading frame 55 (C20orf55), transcript variant 1, mRNA.
USP7	-0.16	0.0000451	0.000904	16	Homo sapiens ubiquitin specific peptidase 7 (herpes virus-associated) (USP7), mRNA.
GCN1L1	-0.11	0.0000482	0.00094	12	Homo sapiens GCN1 general control of amino-acid synthesis 1-like 1 (yeast) (GCN1L1), mRNA.
SRRM2	-0.11	0.0000487	0.000941	16	Homo sapiens serine/arginine repetitive matrix 2 (SRRM2), mRNA.
DIAPH1	-0.15	0.0000488	0.000941	5	Homo sapiens diaphanous homolog 1 (Drosophila) (DIAPH1), transcript variant 1, mRNA.
DNAJB6	-0.11	0.0000495	0.000943	7	Homo sapiens DnaJ (Hsp40) homolog, subfamily B, member 6 (DNAJB6), transcript variant 1, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
CORO7	-0.11	0.0000508	0.000946	16	Homo sapiens coronin 7 (CORO7), mRNA.
CD7	-0.14	0.0000508	0.000946	17	Homo sapiens CD7 molecule (CD7), mRNA.
ABCF1	-0.11	0.000051	0.000946	6	Homo sapiens ATP-binding cassette, sub-family F (GCN20), member 1 (ABCF1), transcript variant 1, mRNA.
POLDIP3	-0.13	0.0000516	0.000953	22	Homo sapiens polymerase (DNA-directed), delta interacting protein 3 (POLDIP3), transcript variant 2, mRNA.
SMARCA4	-0.12	0.0000526	0.000964	19	Homo sapiens SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 (SMARCA4), mRNA.
SEPT5	-0.28	0.0000535	0.000977	22	Homo sapiens septin 5 (SEPT5), mRNA.
UBA1	-0.16	0.0000554	0.00101	X	Homo sapiens ubiquitin-like modifier activating enzyme 1 (UBA1), transcript variant 1, mRNA.
DDX24	-0.12	0.000056	0.00101	14	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 24 (DDX24), mRNA.
RALY	-0.1	0.0000566	0.00101	20	Homo sapiens RNA binding protein, autoantigenic (hnRNP-associated with lethal yellow homolog (mouse)) (RALY), transcript variant 2, mRNA.
RNF31	-0.12	0.0000597	0.00104	14	Homo sapiens ring finger protein 31 (RNF31), mRNA.
NRGN	-0.28	0.0000622	0.00106	11	Homo sapiens neurogranin (protein kinase C substrate, RC3) (NRGN), mRNA.
NCOR2	-0.13	0.0000623	0.00106	12	Homo sapiens nuclear receptor co-repressor 2 (NCOR2), transcript variant 1, mRNA.
PIK3CD	-0.13	0.0000625	0.00106	1	Homo sapiens phosphoinositide-3-kinase, catalytic, delta polypeptide (PIK3CD), mRNA.
HEXDC	-0.11	0.0000628	0.00106	17	Homo sapiens hexosaminidase (glycosyl hydrolase family 20, catalytic domain) containing (HEXDC), mRNA.
UBE1	-0.15	0.0000646	0.00108	X	Homo sapiens ubiquitin-activating enzyme E1 (UBE1), transcript variant 1, mRNA.
DYNC1H1	-0.12	0.0000716	0.00114	14	Homo sapiens dynein, cytoplasmic 1, heavy chain 1 (DYNC1H1), mRNA.
STAG3L2	-0.13	0.0000722	0.00114	7	Homo sapiens stromal antigen 3-like 2 (STAG3L2), mRNA.
DCTN1	-0.13	0.0000725	0.00114	2	Homo sapiens dynactin 1 (p150, glued homolog, Drosophila) (DCTN1), transcript variant 2, mRNA.
HSPB1	-0.15	0.0000731	0.00114	7	Homo sapiens heat shock 27kDa protein 1 (HSPB1), mRNA.
CLPTM1	-0.11	0.0000734	0.00114	19	Homo sapiens cleft lip and palate associated transmembrane protein 1 (CLPTM1), mRNA.
MOBKL2A	-0.16	0.0000734	0.00114	19	Homo sapiens MOB1, Mps One Binder kinase activator-like 2A (yeast) (MOBKL2A), mRNA.
C21ORF58	-0.11	0.0000789	0.0012	21	Homo sapiens chromosome 21 open reading frame 58 (C21orf58), mRNA.
WDR13	-0.1	0.0000809	0.00121	X	Homo sapiens WD repeat domain 13 (WDR13), mRNA.
STAG3L3	-0.13	0.0000846	0.00125	7	Homo sapiens stromal antigen 3-like 3 (STAG3L3), mRNA.
TRIM41	-0.12	0.0000912	0.0013	5	Homo sapiens tripartite motif-containing 41 (TRIM41), transcript variant 2, mRNA.
TAGLN2	-0.15	0.0000914	0.0013	1	Homo sapiens transgelin 2 (TAGLN2), mRNA.
ESYT1	-0.11	0.0000916	0.0013	12	Homo sapiens extended synaptotagmin-like protein 1 (ESYT1), mRNA.
INTS3	-0.12	0.0000928	0.0013	1	Homo sapiens integrator complex subunit 3 (INTS3), mRNA.
PHKA2	-0.1	0.0000963	0.00133	X	Homo sapiens phosphorylase kinase, alpha 2 (liver) (PHKA2), mRNA.
CARD11	-0.11	0.0000984	0.00135	7	Homo sapiens caspase recruitment domain family, member 11 (CARD11), mRNA.
CTNS	-0.1	0.0000993	0.00135	17	Homo sapiens cystinosis, nephropathic (CTNS), transcript variant 2, mRNA.
GPR137	-0.1	0.000102	0.00138	11	Homo sapiens G protein-coupled receptor 137 (GPR137), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
KPNA6	-0.12	0.000103	0.00138	1	Homo sapiens karyopherin alpha 6 (importin alpha 7) (KPNA6), mRNA.
SLC35A4	-0.11	0.000107	0.00141	5	Homo sapiens solute carrier family 35, member A4 (SLC35A4), mRNA.
TNRC6B	-0.1	0.000108	0.00142	22	Homo sapiens trinucleotide repeat containing 6B (TNRC6B), transcript variant 2, mRNA.
DPF2	-0.11	0.000108	0.00142	11	Homo sapiens D4, zinc and double PHD fingers family 2 (DPF2), mRNA.
TCTA	-0.16	0.000113	0.00145	3	Homo sapiens T-cell leukemia translocation altered gene (TCTA), mRNA.
CLDN14	-0.11	0.000115	0.00147	21	Homo sapiens claudin 14 (CLDN14), transcript variant 2, mRNA.
SEC31A	-0.1	0.000119	0.0015	4	Homo sapiens SEC31 homolog A (S. cerevisiae) (SEC31A), transcript variant 1, mRNA.
FLJ12078	-0.12	0.00012	0.0015		PREDICTED: Homo sapiens hypothetical protein FLJ12078 (FLJ12078), misc RNA.
TSPAN33	-0.19	0.000121	0.00151	7	Homo sapiens tetraspanin 33 (TSPAN33), mRNA.
DNM2	-0.11	0.000122	0.00152	19	Homo sapiens dynamin 2 (DNM2), transcript variant 2, mRNA.
AATF	-0.1	0.000132	0.0016	17	Homo sapiens apoptosis antagonizing transcription factor (AATF), mRNA.
SIPA1	-0.1	0.000134	0.00161	11	Homo sapiens signal-induced proliferation-associated gene 1 (SIPA1), transcript variant 2, mRNA.
EIF4G1	-0.1	0.000139	0.00164	3	Homo sapiens eukaryotic translation initiation factor 4 gamma, 1 (EIF4G1), transcript variant 1, mRNA.
SEC16A	-0.11	0.00014	0.00164	9	Homo sapiens SEC16 homolog A (S. cerevisiae) (SEC16A), mRNA.
RHOC	-0.14	0.000142	0.00165	1	Homo sapiens ras homolog gene family, member C (RHOC), transcript variant 1, mRNA.
LAT	-0.11	0.000146	0.00167	16	Homo sapiens linker for activation of T cells (LAT), transcript variant 2, mRNA.
PLOD3	-0.1	0.000149	0.00169	7	Homo sapiens procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3 (PLOD3), mRNA.
SP2	-0.15	0.00016	0.00175	17	Homo sapiens Sp2 transcription factor (SP2), mRNA.
CCDC92	-0.1	0.000175	0.00186	12	Homo sapiens coiled-coil domain containing 92 (CCDC92), mRNA.
ILF3	-0.1	0.000177	0.00187	19	Homo sapiens interleukin enhancer binding factor 3, 90kDa (ILF3), transcript variant 1, mRNA.
SERTAD1	-0.13	0.000182	0.00188	19	Homo sapiens SERTA domain containing 1 (SERTAD1), mRNA.
PIAS4	-0.11	0.000185	0.0019	19	Homo sapiens protein inhibitor of activated STAT, 4 (PIAS4), mRNA.
SLC44A2	-0.1	0.000188	0.00193	19	Homo sapiens solute carrier family 44, member 2 (SLC44A2), mRNA.
SEMA4D	-0.1	0.000191	0.00195	9	Homo sapiens sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4D (SEMA4D), mRNA.
RNF40	-0.16	0.000194	0.00196	16	Homo sapiens ring finger protein 40 (RNF40), mRNA.
FOXJ2	-0.11	0.000198	0.00199	12	Homo sapiens forkhead box J2 (FOXJ2), mRNA.
GATAD2A	-0.1	0.000202	0.002	19	Homo sapiens GATA zinc finger domain containing 2A (GATAD2A), mRNA.
PNPLA6	-0.1	0.000205	0.00202	19	Homo sapiens patatin-like phospholipase domain containing 6 (PNPLA6), mRNA.
FAM53B	-0.11	0.000206	0.00202	10	Homo sapiens family with sequence similarity 53, member B (FAM53B), mRNA.
POLR2A	-0.14	0.000229	0.00214	17	Homo sapiens polymerase (RNA) II (DNA directed) polypeptide A, 220kDa (POLR2A), mRNA.
RBM6	-0.1	0.000232	0.00216	3	Homo sapiens RNA binding motif protein 6 (RBM6), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
PAF1	-0.11	0.000237	0.00218	19	Homo sapiens Paf1, RNA polymerase II associated factor, homolog (S. cerevisiae) (PAF1), mRNA.
SBK1	-0.12	0.000253	0.00229	16	Homo sapiens SH3-binding domain kinase 1 (SBK1), mRNA.
PSD4	-0.12	0.00027	0.00237	2	Homo sapiens pleckstrin and Sec7 domain containing 4 (PSD4), mRNA.
IRF3	-0.23	0.000284	0.00245	19	Homo sapiens interferon regulatory factor 3 (IRF3), mRNA.
TBCD	-0.13	0.000292	0.00249	17	Homo sapiens tubulin folding cofactor D (TBCD), mRNA.
MLL4	-0.11	0.000294	0.0025	19	Homo sapiens myeloid/lymphoid or mixed-lineage leukemia 4 (MLL4), mRNA.
IRX1	-0.28	0.000296	0.00251	5	Homo sapiens iroquois homeobox 1 (IRX1), mRNA.
SH2D3C	-0.12	0.000307	0.00258	9	Homo sapiens SH2 domain containing 3C (SH2D3C), transcript variant 2, mRNA.
INPPL1	-0.12	0.000311	0.00261	11	Homo sapiens inositol polyphosphate phosphatase-like 1 (INPPL1), mRNA.
UBE2L3	-0.12	0.000328	0.00272	22	Homo sapiens ubiquitin-conjugating enzyme E2L 3 (UBE2L3), transcript variant 2, mRNA.
CSNK2A1	-0.23	0.000343	0.0028	20	Homo sapiens casein kinase 2, alpha 1 polypeptide (CSNK2A1), transcript variant 2, mRNA.
STK10	-0.11	0.000344	0.0028	5	Homo sapiens serine/threonine kinase 10 (STK10), mRNA.
SMARCC2	-0.13	0.000346	0.0028	12	Homo sapiens SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2 (SMARCC2), transcript variant 2, mRNA.
SIDT2	-0.15	0.000348	0.00281	11	Homo sapiens SID1 transmembrane family, member 2 (SIDT2), mRNA.
ZNF493	-0.11	0.000349	0.00281	19	Homo sapiens zinc finger protein 493 (ZNF493), transcript variant 3, mRNA.
FAM38A	-0.11	0.00036	0.00286	16	Homo sapiens family with sequence similarity 38, member A (FAM38A), mRNA.
ITGAL	-0.1	0.000364	0.00288	16	Homo sapiens integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide) (ITGAL), mRNA.
BAT3	-0.16	0.000367	0.0029	6	Homo sapiens HLA-B associated transcript 3 (BAT3), transcript variant 3, mRNA.
CASC3	-0.12	0.000373	0.00292	17	Homo sapiens cancer susceptibility candidate 3 (CASC3), mRNA.
CDKN1A	-0.18	0.000374	0.00293	6	Homo sapiens cyclin-dependent kinase inhibitor 1A (p21, Cip1) (CDKN1A), transcript variant 1, mRNA.
LPAR5	-0.11	0.000384	0.00299	12	Homo sapiens lysophosphatidic acid receptor 5 (LPAR5), mRNA.
FKBP8	-0.26	0.000415	0.00312	19	Homo sapiens FK506 binding protein 8, 38kDa (FKBP8), mRNA.
MCM5	-0.1	0.000421	0.00312	22	Homo sapiens minichromosome maintenance complex component 5 (MCM5), mRNA.
FBXO18	-0.1	0.000437	0.00317	10	Homo sapiens F-box protein, helicase, 18 (FBXO18), transcript variant 2, mRNA.
ITPK1	-0.14	0.000437	0.00317	14	Homo sapiens inositol 1,3,4-triphosphate 5/6 kinase (ITPK1), mRNA.
UBN1	-0.14	0.000454	0.00326	16	Homo sapiens ubinuclein 1 (UBN1), transcript variant 2, mRNA.
PTPN7	-0.13	0.000458	0.00326	1	Homo sapiens protein tyrosine phosphatase, non-receptor type 7 (PTPN7), transcript variant 2, mRNA.
VPS37B	-0.1	0.000476	0.00337	12	Homo sapiens vacuolar protein sorting 37 homolog B (S. cerevisiae) (VPS37B), mRNA.
AKNA	-0.1	0.000481	0.00339		Homo sapiens AT-hook transcription factor (AKNA), mRNA.
TAF15	-0.13	0.00049	0.00342	17	Homo sapiens TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa (TAF15), transcript variant 1, mRNA.
SMG7	-0.1	0.000502	0.00345	1	Homo sapiens Smg-7 homolog, nonsense mediated mRNA decay factor (C. elegans) (SMG7), transcript variant 1, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
VAV1	-0.11	0.000504	0.00346	19	Homo sapiens vav 1 guanine nucleotide exchange factor (VAV1), mRNA.
MYO9B	-0.1	0.00051	0.00348	19	Homo sapiens myosin IXB (MYO9B), mRNA.
MLLT6	-0.1	0.000522	0.00352	17	Homo sapiens myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 6 (MLLT6), mRNA.
CLIC3	-0.23	0.000525	0.00353	9	Homo sapiens chloride intracellular channel 3 (CLIC3), mRNA.
ARHGAP1	-0.1	0.000525	0.00353	11	Homo sapiens Rho GTPase activating protein 1 (ARHGAP1), mRNA.
CMIP	-0.12	0.000532	0.00357	16	Homo sapiens c-Maf-inducing protein (CMIP), transcript variant Tc-mip, mRNA.
SEMA4B	-0.11	0.000546	0.00362	15	Homo sapiens sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4B (SEMA4B), transcript variant 2, mRNA.
VCL	-0.13	0.000552	0.00363	10	Homo sapiens vinculin (VCL), transcript variant 1, mRNA.
MARCKSL1	-0.1	0.000588	0.00377	1	Homo sapiens MARCKS-like 1 (MARCKSL1), mRNA.
MYADM	-0.15	0.000596	0.00379	19	Homo sapiens myeloid-associated differentiation marker (MYADM), transcript variant 4, mRNA.
ACTRT1	-0.29	0.000598	0.0038	X	Homo sapiens actin-related protein T1 (ACTRT1), mRNA.
CABIN1	-0.11	0.000602	0.00382	22	Homo sapiens calcineurin binding protein 1 (CABIN1), mRNA.
MBD6	-0.12	0.000606	0.00382	12	Homo sapiens methyl-CpG binding domain protein 6 (MBD6), mRNA.
ARF3	-0.12	0.000608	0.00382	12	Homo sapiens ADP-ribosylation factor 3 (ARF3), mRNA.
FAM153B	-0.18	0.00062	0.00387	5	Homo sapiens family with sequence similarity 153, member B (FAM153B), mRNA.
VWCE	-0.37	0.000639	0.00394	11	Homo sapiens von Willebrand factor C and EGF domains (VWCE), mRNA.
HCFC1	-0.12	0.00071	0.00424	X	Homo sapiens host cell factor C1 (VP16-accessory protein) (HCFC1), mRNA.
FBRS	-0.11	0.000712	0.00425	16	Homo sapiens fibrosin (FBRS), mRNA.
DHX34	-0.13	0.00073	0.00433	19	Homo sapiens DEAH (Asp-Glu-Ala-His) box polypeptide 34 (DHX34), mRNA.
ITGA5	-0.11	0.000748	0.00438	12	Homo sapiens integrin, alpha 5 (fibronectin receptor, alpha polypeptide) (ITGA5), mRNA.
ZYG11B	-0.11	0.000783	0.00451	1	Homo sapiens zyg-11 homolog B (C. elegans) (ZYG11B), mRNA.
TPP1	-0.13	0.000785	0.00451	11	Homo sapiens tripeptidyl peptidase I (TPP1), mRNA.
UBXN6	-0.22	0.000785	0.00451	19	Homo sapiens UBX domain protein 6 (UBXN6), mRNA.
TRIM38	-0.11	0.000794	0.00455	6	Homo sapiens tripartite motif-containing 38 (TRIM38), mRNA.
CYBASC3	-0.1	0.000797	0.00456	11	Homo sapiens cytochrome b, ascorbate dependent 3 (CYBASC3), mRNA.
TRIM26	-0.1	0.00081	0.0046	6	Homo sapiens tripartite motif-containing 26 (TRIM26), mRNA.
ARAF	-0.14	0.000816	0.00461	X	Homo sapiens v-raf murine sarcoma 3611 viral oncogene homolog (ARAF), mRNA.
MBNL3	-0.26	0.000834	0.00467	X	Homo sapiens muscleblind-like 3 (Drosophila) (MBNL3), transcript variant R, mRNA.
CARM1	-0.19	0.000846	0.00471	19	Homo sapiens coactivator-associated arginine methyltransferase 1 (CARM1), mRNA.
PPP2R2B	-0.1	0.000857	0.00475	5	Homo sapiens protein phosphatase 2 (formerly 2A), regulatory subunit B, beta isoform (PPP2R2B), transcript variant 5, mRNA.
DTX2	-0.1	0.000877	0.00481	7	Homo sapiens deltex homolog 2 (Drosophila) (DTX2), mRNA.
TSC22D1	-0.16	0.000885	0.00483	13	Homo sapiens TSC22 domain family, member 1 (TSC22D1), transcript variant 2, mRNA.
PILRB	-0.11	0.000885	0.00483	7	Homo sapiens paired immunoglobin-like type 2 receptor beta (PILRB), transcript variant 1, mRNA.
ANKRD30B	-0.1	0.000897	0.00484	18	Homo sapiens ankyrin repeat domain 30B (ANKRD30B), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
LRRC33	-0.1	0.00092	0.00489	3	Homo sapiens leucine rich repeat containing 33 (LRRC33), mRNA.
RRBP1	-0.11	0.000921	0.00489	20	Homo sapiens ribosome binding protein 1 homolog 180kDa (dog) (RRBP1), transcript variant 1, mRNA.
CD37	-0.1	0.000926	0.0049	19	Homo sapiens CD37 antigen (CD37), mRNA.
NUAK2	-0.13	0.000927	0.0049	1	Homo sapiens NUAK family, SNF1-like kinase, 2 (NUAK2), mRNA.
TARDBP	-0.16	0.000947	0.00494	1	Homo sapiens TAR DNA binding protein (TARDBP), mRNA.
MSN	-0.13	0.00095	0.00495	X	Homo sapiens moesin (MSN), mRNA.
LYL1	-0.17	0.000957	0.00498	19	Homo sapiens lymphoblastic leukemia derived sequence 1 (LYL1), mRNA.
CLK2	-0.25	0.000961	0.00499		PREDICTED: Homo sapiens CDC-like kinase 2, transcript variant 4 (CLK2), mRNA.
TUBB1	-0.2	0.000971	0.00501	20	Homo sapiens tubulin, beta 1 (TUBB1), mRNA.
CKAP5	-0.1	0.000984	0.00506	11	Homo sapiens cytoskeleton associated protein 5 (CKAP5), transcript variant 1, mRNA.
NOMO2	-0.1	0.001	0.00514	16	Homo sapiens NODAL modulator 2 (NOMO2), transcript variant 2, mRNA.
PNPLA7	-0.15	0.00102	0.00517	9	Homo sapiens patatin-like phospholipase domain containing 7 (PNPLA7), mRNA.
MYH9	-0.14	0.00102	0.00519	22	Homo sapiens myosin, heavy chain 9, non-muscle (MYH9), mRNA.
PXN	-0.11	0.00105	0.00527	12	Homo sapiens paxillin (PXN), mRNA.
TMEM140	-0.15	0.00105	0.00529	7	Homo sapiens transmembrane protein 140 (TMEM140), mRNA.
C9ORF164	-0.12	0.00106	0.00532	9	Homo sapiens chromosome 9 open reading frame 164 (C9orf164), mRNA.
ADRA2C	-0.29	0.00107	0.00533		PREDICTED: Homo sapiens adrenergic, alpha-2C-, receptor (ADRA2C), mRNA.
CXXC5	-0.12	0.00107	0.00536	5	Homo sapiens CXXC finger 5 (CXXC5), mRNA.
CTSB	-0.13	0.00108	0.00537	8	Homo sapiens cathepsin B (CTSB), transcript variant 1, mRNA.
CSF1R	-0.13	0.00108	0.00538	5	Homo sapiens colony stimulating factor 1 receptor, formerly McDonough feline sarcoma viral (v-fms) oncogene homolog (CSF1R), mRNA.
SLC6A10P	-0.31	0.00109	0.00542	16	Homo sapiens solute carrier family 6 (neurotransmitter transporter, creatine), member 10 (pseudogene) (SLC6A10P) on chromosome 16.
DDR1	-0.26	0.0011	0.00542	6	Homo sapiens discoidin domain receptor tyrosine kinase 1 (DDR1), transcript variant 1, mRNA.
WBP1	-0.11	0.00113	0.00549	2	Homo sapiens WW domain binding protein 1 (WBP1), mRNA.
GMIP	-0.11	0.00114	0.00551	19	Homo sapiens GEM interacting protein (GMIP), mRNA.
MEGF10	-0.35	0.00114	0.00552	5	Homo sapiens multiple EGF-like-domains 10 (MEGF10), mRNA.
HLA-DRB6	-0.1	0.00118	0.00564	6	Homo sapiens major histocompatibility complex, class II, DR beta 6 (pseudogene) (HLA-DRB6), non-coding RNA.
PVRIG	-0.11	0.00118	0.00564	7	Homo sapiens poliovirus receptor related immunoglobulin domain containing (PVRIG), mRNA.
PRPF8	-0.13	0.00121	0.00575	17	Homo sapiens PRP8 pre-mRNA processing factor 8 homolog (S. cerevisiae) (PRPF8), mRNA.
SNAPC2	-0.15	0.00124	0.00584	19	Homo sapiens small nuclear RNA activating complex, polypeptide 2, 45kDa (SNAPC2), mRNA.
COQ6	-0.16	0.00124	0.00585	14	Homo sapiens coenzyme Q6 homolog, monooxygenase (S. cerevisiae) (COQ6), transcript variant 1, mRNA.
TOP3A	-0.12	0.00125	0.00586	17	Homo sapiens topoisomerase (DNA) III alpha (TOP3A), mRNA.
ERGIC1	-0.12	0.00126	0.00591	5	Homo sapiens endoplasmic reticulum-golgi intermediate compartment (ERGIC) 1 (ERGIC1), transcript variant 1, mRNA.
MARCH2	-0.17	0.00127	0.00593	19	Homo sapiens membrane-associated ring finger (C3HC4) 2 (MARCH2), transcript variant 3, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
STK40	-0.1	0.00127	0.00593	1	Homo sapiens serine/threonine kinase 40 (STK40), mRNA.
MKNK2	-0.1	0.00127	0.00594	19	Homo sapiens MAP kinase interacting serine/threonine kinase 2 (MKNK2), transcript variant 1, mRNA.
FAM65A	-0.1	0.00129	0.00599	16	Homo sapiens family with sequence similarity 65, member A (FAM65A), mRNA.
FCGR3A	-0.35	0.0013	0.006	1	Homo sapiens Fc fragment of IgG, low affinity IIIa, receptor (CD16a) (FCGR3A), transcript variant 1, mRNA.
SEC14L1	-0.15	0.0013	0.006	17	Homo sapiens SEC14-like 1 (S. cerevisiae) (SEC14L1), transcript variant 1, mRNA.
CENTD2	-0.11	0.00133	0.00612	11	Homo sapiens centaurin, delta 2 (CENTD2), transcript variant 3, mRNA.
CTSA	-0.13	0.00133	0.00615	20	Homo sapiens cathepsin A (CTSA), transcript variant 1, mRNA.
EIF2C2	-0.14	0.00137	0.00629	8	Homo sapiens eukaryotic translation initiation factor 2C, 2 (EIF2C2), mRNA.
FZD7	-0.23	0.00137	0.00629	2	Homo sapiens frizzled homolog 7 (Drosophila) (FZD7), mRNA.
RUNX3	-0.1	0.00138	0.00631	1	Homo sapiens runt-related transcription factor 3 (RUNX3), transcript variant 2, mRNA.
PTGS1	-0.19	0.00142	0.00643	9	Homo sapiens prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase) (PTGS1), transcript variant 2, mRNA.
USP4	-0.1	0.00143	0.00646	3	Homo sapiens ubiquitin specific peptidase 4 (proto-oncogene) (USP4), transcript variant 1, mRNA.
CPSF1	-0.11	0.00144	0.00647	8	Homo sapiens cleavage and polyadenylation specific factor 1, 160kDa (CPSF1), mRNA.
LRCH4	-0.1	0.00145	0.00651	7	Homo sapiens leucine-rich repeats and calponin homology (CH) domain containing 4 (LRCH4), mRNA.
CHD4	-0.19	0.0015	0.00668	12	Homo sapiens chromodomain helicase DNA binding protein 4 (CHD4), mRNA.
SF3A2	-0.1	0.00151	0.0067	19	Homo sapiens splicing factor 3a, subunit 2, 66kDa (SF3A2), mRNA.
ZAP70	-0.1	0.00153	0.00677	2	Homo sapiens zeta-chain (TCR) associated protein kinase 70kDa (ZAP70), transcript variant 1, mRNA.
ALDOC	-0.11	0.00157	0.00689	17	Homo sapiens aldolase C, fructose-bisphosphate (ALDOC), mRNA.
C2ORF24	-0.16	0.00159	0.00695	2	Homo sapiens chromosome 2 open reading frame 24 (C2orf24), mRNA.
HDAC6	-0.22	0.00162	0.00701	X	Homo sapiens histone deacetylase 6 (HDAC6), mRNA.
C7ORF41	-0.13	0.00167	0.00718	7	Homo sapiens chromosome 7 open reading frame 41 (C7orf41), mRNA.
TGFBR3	-0.16	0.00167	0.00718	1	Homo sapiens transforming growth factor, beta receptor III (TGFBR3), mRNA.
CREBBP	-0.11	0.00168	0.00721	16	Homo sapiens CREB binding protein (CREBBP), transcript variant 2, mRNA.
ATP6V0C	-0.14	0.00171	0.00729		PREDICTED: Homo sapiens ATPase, H+ transporting, lysosomal 16kDa, V0 subunit c (ATP6V0C), mRNA.
FLJ20699	-0.16	0.00184	0.00769	22	Homo sapiens hypothetical protein FLJ20699 (FLJ20699), mRNA.
CHD8	-0.1	0.00187	0.00777	14	Homo sapiens chromodomain helicase DNA binding protein 8 (CHD8), mRNA.
DNA2	-0.28	0.00187	0.00777	10	Homo sapiens DNA replication helicase 2 homolog (yeast) (DNA2), mRNA.
LRP10	-0.1	0.00194	0.00795	14	Homo sapiens low density lipoprotein receptor-related protein 10 (LRP10), mRNA.
DENR	-0.12	0.00197	0.00802	12	Homo sapiens density-regulated protein (DENR), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
CD79A	-0.15	0.00199	0.00805	19	Homo sapiens CD79a molecule, immunoglobulin-associated alpha (CD79A), transcript variant 2, mRNA.
FOXO4	-0.15	0.002	0.0081	X	Homo sapiens forkhead box O4 (FOXO4), mRNA.
IGF2R	-0.13	0.00205	0.00826	6	Homo sapiens insulin-like growth factor 2 receptor (IGF2R), mRNA.
PRR5	-0.13	0.00209	0.00838	22	Homo sapiens proline rich 5 (renal) (PRR5), transcript variant 5, mRNA.
CAPN1	-0.1	0.0021	0.0084	11	Homo sapiens calpain 1, (mu/I) large subunit (CAPN1), mRNA.
TIMP1	-0.11	0.00211	0.00843	X	Homo sapiens TIMP metallopeptidase inhibitor 1 (TIMP1), mRNA.
PI4KAP2	-0.1	0.00212	0.00844		Homo sapiens phosphatidylinositol 4-kinase, catalytic, alpha polypeptide pseudogene 2 (PI4KAP2), mRNA.
FCRL2	-0.25	0.00214	0.00851	1	Homo sapiens Fc receptor-like 2 (FCRL2), transcript variant 2, mRNA.
PPP1R15A	-0.1	0.0022	0.00867	19	Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 15A (PPP1R15A), mRNA.
SORL1	-0.13	0.00224	0.00878	11	Homo sapiens sortilin-related receptor, L(DLR class) A repeats-containing (SORL1), mRNA.
GPR175	-0.22	0.00226	0.0088	3	Homo sapiens G protein-coupled receptor 175 (GPR175), mRNA.
SPOCK2	-0.1	0.00227	0.00882	10	Homo sapiens sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2 (SPOCK2), mRNA.
CD3E	-0.11	0.0023	0.00891	11	Homo sapiens CD3e molecule, epsilon (CD3-TCR complex) (CD3E), mRNA.
ZNFX1	-0.11	0.0023	0.00891	20	Homo sapiens zinc finger, NFX1-type containing 1 (ZNFX1), mRNA.
ATG2A	-0.12	0.0023	0.00891	11	Homo sapiens ATG2 autophagy related 2 homolog A (S. cerevisiae) (ATG2A), mRNA.
MAP1S	-0.16	0.00232	0.00898	19	Homo sapiens microtubule-associated protein 1S (MAP1S), mRNA.
BRD3	-0.12	0.00239	0.00917	9	Homo sapiens bromodomain containing 3 (BRD3), mRNA.
ATXN10	-0.26	0.00239	0.00917	22	Homo sapiens ataxin 10 (ATXN10), mRNA.
PDCD4	-0.11	0.00242	0.00926	10	Homo sapiens programmed cell death 4 (neoplastic transformation inhibitor) (PDCD4), transcript variant 2, mRNA.
MAST3	-0.11	0.00255	0.00956	19	Homo sapiens microtubule associated serine/threonine kinase 3 (MAST3), mRNA.
ACTN4	-0.13	0.00257	0.00964	19	Homo sapiens actinin, alpha 4 (ACTN4), mRNA.
RNF10	-0.23	0.00258	0.00965	12	Homo sapiens ring finger protein 10 (RNF10), mRNA.
FXR2	-0.17	0.00258	0.00966	17	Homo sapiens fragile X mental retardation, autosomal homolog 2 (FXR2), mRNA.
EIF2AK4	-0.12	0.00265	0.00982	15	Homo sapiens eukaryotic translation initiation factor 2 alpha kinase 4 (EIF2AK4), mRNA.
TAPBP	-0.1	0.00268	0.00989	6	Homo sapiens TAP binding protein (tapasin) (TAPBP), transcript variant 1, mRNA.
MIDN	-0.11	0.00271	0.00994	19	Homo sapiens midnolin (MIDN), mRNA.
SNTB2	-0.1	0.00275	0.0101	16	Homo sapiens syntrophin, beta 2 (dystrophin-associated protein A1, 59kDa, basic component 2) (SNTB2), transcript variant 1, mRNA.
ELMO1	-0.1	0.00289	0.0105	7	Homo sapiens engulfment and cell motility 1 (ELMO1), transcript variant 1, mRNA.
MAP7D1	-0.12	0.0029	0.0105	1	Homo sapiens MAP7 domain containing 1 (MAP7D1), mRNA.
STXBP2	-0.11	0.00299	0.0108	19	Homo sapiens syntaxin binding protein 2 (STXBP2), mRNA.
PNPLA2	-0.14	0.00313	0.0111	11	Homo sapiens patatin-like phospholipase domain containing 2 (PNPLA2), mRNA.
FLJ35390	-0.12	0.00318	0.0113	7	Homo sapiens hypothetical LOC255031 (FLJ35390), transcript variant 1, non-coding RNA.
IL8RB	-0.14	0.00321	0.0113	2	Homo sapiens interleukin 8 receptor, beta (IL8RB), mRNA.
C22ORF25	-0.13	0.00321	0.0113	22	Homo sapiens chromosome 22 open reading frame 25 (C22orf25), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
S1PR5	-0.18	0.00325	0.0115	19	Homo sapiens sphingosine-1-phosphate receptor 5 (S1PR5), mRNA.
SPTAN1	-0.11	0.00335	0.0117	9	Homo sapiens spectrin, alpha, non-erythrocytic 1 (alpha-fodrin) (SPTAN1), mRNA.
N4BP2	-0.11	0.00343	0.0119	4	Homo sapiens Nedd4 binding protein 2 (N4BP2), mRNA.
MOSPD3	-0.22	0.00343	0.0119	7	Homo sapiens motile sperm domain containing 3 (MOSPD3), transcript variant 2, mRNA.
PYGB	-0.1	0.0035	0.0121	20	Homo sapiens phosphorylase, glycogen; brain (PYGB), mRNA.
FLJ38717	-0.1	0.00359	0.0123	6	Homo sapiens FLJ38717 protein (FLJ38717), mRNA.
TNFSF14	-0.11	0.0036	0.0123	19	Homo sapiens tumor necrosis factor (ligand) superfamily, member 14 (TNFSF14), transcript variant 2, mRNA.
MAP4K4	-0.2	0.00361	0.0124	2	Homo sapiens mitogen-activated protein kinase kinase kinase kinase 4 (MAP4K4), transcript variant 3, mRNA.
DOPEY2	-0.12	0.00364	0.0124	21	Homo sapiens dopey family member 2 (DOPEY2), mRNA.
DPYSL5	-0.27	0.00366	0.0125	2	Homo sapiens dihydropyrimidinase-like 5 (DPYSL5), mRNA.
SIPA1L3	-0.23	0.00384	0.013	19	Homo sapiens signal-induced proliferation-associated 1 like 3 (SIPA1L3), mRNA.
SLC9A1	-0.1	0.00389	0.0131	1	Homo sapiens solute carrier family 9 (sodium/hydrogen exchanger), member 1 (SLC9A1), mRNA.
LYPD3	-0.25	0.00395	0.0133	19	Homo sapiens LY6/PLAUR domain containing 3 (LYPD3), mRNA.
SYTL3	-0.1	0.00402	0.0134	6	Homo sapiens synaptotagmin-like 3 (SYTL3), mRNA.
ST6GALNAC4	-0.2	0.00407	0.0136	9	Homo sapiens ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1, 3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 4 (ST6GALNAC4), transcript variant 1, mRNA.
FASN	-0.12	0.00408	0.0136	17	Homo sapiens fatty acid synthase (FASN), mRNA.
UBE2W	-0.22	0.00409	0.0136	8	Homo sapiens ubiquitin-conjugating enzyme E2W (putative) (UBE2W), transcript variant 2, mRNA.
ZNF669	-0.1	0.00414	0.0137	1	Homo sapiens zinc finger protein 669 (ZNF669), mRNA.
HDAC7A	-0.24	0.00417	0.0137	12	Homo sapiens histone deacetylase 7A (HDAC7A), transcript variant 2, mRNA.
SMU1	-0.19	0.00426	0.014	9	Homo sapiens smu-1 suppressor of mec-8 and unc-52 homolog (C. elegans) (SMU1), mRNA.
ZNF786	-0.1	0.00438	0.0143	7	Homo sapiens zinc finger protein 786 (ZNF786), mRNA.
MT2A	-0.17	0.00443	0.0144	16	Homo sapiens metallothionein 2A (MT2A), mRNA.
TTC38	-0.15	0.00448	0.0146	22	Homo sapiens tetratricopeptide repeat domain 38 (TTC38), mRNA.
LSP1	-0.12	0.0045	0.0146	11	Homo sapiens lymphocyte-specific protein 1 (LSP1), transcript variant 2, mRNA.
RARA	-0.1	0.00454	0.0147	17	Homo sapiens retinoic acid receptor, alpha (RARA), transcript variant 1, mRNA.
ANKRD13D	-0.13	0.00456	0.0148		PREDICTED: Homo sapiens ankyrin repeat domain 13 family, member D, transcript variant 7 (ANKRD13D), mRNA.
SLC25A39	-0.15	0.00465	0.015	17	Homo sapiens solute carrier family 25, member 39 (SLC25A39), mRNA.
NOC4L	-0.22	0.0047	0.0151		PREDICTED: Homo sapiens nucleolar complex associated 4 homolog (S. cerevisiae) (NOC4L), mRNA.
RNF24	-0.12	0.0047	0.0151	20	Homo sapiens ring finger protein 24 (RNF24), mRNA.
AHR	-0.1	0.0047	0.0151	7	Homo sapiens aryl hydrocarbon receptor (AHR), mRNA.
HIST2H2BE	-0.13	0.00473	0.0151	1	Homo sapiens histone cluster 2, H2be (HIST2H2BE), mRNA.
SPRYD3	-0.24	0.00477	0.0152	12	Homo sapiens SPRY domain containing 3 (SPRYD3), mRNA.
COL8A2	-0.22	0.00483	0.0153	1	Homo sapiens collagen, type VIII, alpha 2 (COL8A2), mRNA.
CTSD	-0.13	0.00498	0.0157	11	Homo sapiens cathepsin D (CTSD), mRNA.
YPEL3	-0.11	0.00508	0.0159	16	Homo sapiens yippee-like 3 (Drosophila) (YPEL3), transcript variant 1, mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
FAM40B	-0.14	0.0051	0.0159	7	Homo sapiens family with sequence similarity 40, member B (FAM40B), mRNA.
C22ORF13	-0.13	0.0051	0.0159	22	Homo sapiens chromosome 22 open reading frame 13 (C22orf13), mRNA.
RERE	-0.1	0.00513	0.016	1	Homo sapiens arginine-glutamic acid dipeptide (RE) repeats (RERE), transcript variant 3, mRNA.
MED1	-0.23	0.00514	0.016	17	Homo sapiens mediator complex subunit 1 (MED1), mRNA.
CEP27	-0.13	0.00518	0.0161	15	Homo sapiens centrosomal protein 27kDa (CEP27), mRNA.
ROPN1B	-0.23	0.00518	0.0161	3	Homo sapiens ropporin, rhophilin associated protein 1B (ROPN1B), mRNA.
DCLRE1C	-0.11	0.00529	0.0164	10	Homo sapiens DNA cross-link repair 1C (PSO2 homolog, S. cerevisiae) (DCLRE1C), transcript variant b, mRNA.
SLC25A45	-0.21	0.00537	0.0166	11	Homo sapiens solute carrier family 25, member 45 (SLC25A45), transcript variant 2, mRNA.
GNL3L	-0.13	0.00542	0.0167	X	Homo sapiens guanine nucleotide binding protein-like 3 (nucleolar)-like (GNL3L), mRNA.
C14ORF173	-0.13	0.0055	0.0169	14	Homo sapiens chromosome 14 open reading frame 173 (C14orf173), transcript variant 2, mRNA.
CTSZ	-0.11	0.00558	0.0171	20	Homo sapiens cathepsin Z (CTSZ), mRNA.
C16ORF35	-0.18	0.00562	0.0172	16	Homo sapiens chromosome 16 open reading frame 35 (C16orf35), transcript variant 2, mRNA.
RUNDC2C	-0.12	0.00575	0.0175	16	Homo sapiens RUN domain containing 2C (RUNDC2C), non-coding RNA.
CD79B	-0.13	0.00575	0.0175	17	Homo sapiens CD79b molecule, immunoglobulin-associated beta (CD79B), transcript variant 3, mRNA.
EOMES	-0.13	0.00575	0.0175	3	Homo sapiens eomesodermin homolog (Xenopus laevis) (EOMES), mRNA.
KIAA0513	-0.1	0.0058	0.0176	16	Homo sapiens KIAA0513 (KIAA0513), mRNA.
TMEM86B	-0.16	0.00588	0.0178	19	Homo sapiens transmembrane protein 86B (TMEM86B), mRNA.
NOL10	-0.21	0.00596	0.018	2	Homo sapiens nucleolar protein 10 (NOL10), mRNA.
RAXL1	-0.11	0.00597	0.018	19	Homo sapiens retina and anterior neural fold homeobox like 1 (RAXL1), mRNA.
NADK	-0.1	0.00601	0.0181	1	Homo sapiens NAD kinase (NADK), mRNA.
FAM134A	-0.1	0.00603	0.0181	2	Homo sapiens family with sequence similarity 134, member A (FAM134A), mRNA.
UBL7	-0.1	0.00607	0.0182	15	Homo sapiens ubiquitin-like 7 (bone marrow stromal cell-derived) (UBL7), transcript variant 2, mRNA.
PKNOX1	-0.15	0.0061	0.0182	21	Homo sapiens PBX/knotted 1 homeobox 1 (PKNOX1), mRNA.
C19ORF22	-0.14	0.00622	0.0185	19	Homo sapiens chromosome 19 open reading frame 22 (C19orf22), mRNA.
LMNA	-0.25	0.00628	0.0186	1	Homo sapiens lamin A/C (LMNA), transcript variant 2, mRNA.
PSENEN	-0.22	0.00639	0.0188	19	Homo sapiens presenilin enhancer 2 homolog (C. elegans) (PSENEN), mRNA.
GYPC	-0.14	0.00643	0.0189	2	Homo sapiens glycophorin C (Gerbich blood group) (GYPC), transcript variant 2, mRNA.
XKR8	-0.1	0.00646	0.019	1	Homo sapiens XK, Kell blood group complex subunit-related family, member 8 (XKR8), mRNA.
ATHL1	-0.14	0.00664	0.0194	11	Homo sapiens ATH1, acid trehalase-like 1 (yeast) (ATHL1), mRNA.
CSDA	-0.16	0.00665	0.0194	12	Homo sapiens cold shock domain protein A (CSDA), mRNA.
VRK3	-0.1	0.00667	0.0194	19	Homo sapiens vaccinia related kinase 3 (VRK3), transcript variant 2, mRNA.
VASP	-0.12	0.00677	0.0197	19	Homo sapiens vasodilator-stimulated phosphoprotein (VASP), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
STX10	-0.12	0.00693	0.02	19	Homo sapiens syntaxin 10 (STX10), mRNA.
TBC1D10B	-0.19	0.00704	0.0202	16	Homo sapiens TBC1 domain family, member 10B (TBC1D10B), mRNA.
CCL5	-0.12	0.00718	0.0205	17	Homo sapiens chemokine (C-C motif) ligand 5 (CCL5), mRNA.
SH3BP1	-0.1	0.00725	0.0207	22	Homo sapiens SH3-domain binding protein 1 (SH3BP1), mRNA.
SLC4A5	-0.13	0.00746	0.0212	2	Homo sapiens solute carrier family 4, sodium bicarbonate cotransporter, member 5 (SLC4A5), transcript variant c, mRNA.
C4ORF34	-0.12	0.00755	0.0214	4	Homo sapiens chromosome 4 open reading frame 34 (C4orf34), mRNA.
IP6K1	-0.1	0.00762	0.0216	3	Homo sapiens inositol hexakisphosphate kinase 1 (IP6K1), transcript variant 2, mRNA.
RNF213	-0.24	0.00764	0.0216	17	Homo sapiens ring finger protein 213 (RNF213), mRNA.
NDE1	-0.1	0.00769	0.0217	16	Homo sapiens nudE nuclear distribution gene E homolog 1 (A. nidulans) (NDE1), mRNA.
ACTN1	-0.11	0.00776	0.0218	14	Homo sapiens actinin, alpha 1 (ACTN1), mRNA.
MAP2K3	-0.15	0.00791	0.0222	17	Homo sapiens mitogen-activated protein kinase kinase 3 (MAP2K3), transcript variant A, mRNA.
APBB1IP	-0.1	0.00802	0.0224	10	Homo sapiens amyloid beta (A4) precursor protein-binding, family B, member 1 interacting protein (APBB1IP), mRNA.
PIP5K2B	-0.1	0.00827	0.023	17	Homo sapiens phosphatidylinositol-4-phosphate 5-kinase, type II, beta (PIP5K2B), transcript variant 2, mRNA.
IGF2BP2	-0.2	0.0085	0.0235	3	Homo sapiens insulin-like growth factor 2 mRNA binding protein 2 (IGF2BP2), transcript variant 1, mRNA.
LLPH	-0.1	0.00877	0.0241	12	Homo sapiens LLP homolog, long-term synaptic facilitation (Aplysia) (LLPH), mRNA.
RTF1	-0.18	0.00895	0.0244	15	Homo sapiens Rtf1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae) (RTF1), mRNA.
IL2RB	-0.12	0.00897	0.0245	22	Homo sapiens interleukin 2 receptor, beta (IL2RB), mRNA.
PATE2	-0.11	0.00919	0.025	11	Homo sapiens prostate and testis expressed 2 (PATE2), mRNA.
STAT3	-0.1	0.00931	0.0252	17	Homo sapiens signal transducer and activator of transcription 3 (acute-phase response factor) (STAT3), transcript variant 3, mRNA.
EPB49	-0.21	0.00935	0.0253	8	Homo sapiens erythrocyte membrane protein band 4.9 (dematin) (EPB49), mRNA.
G6PD	-0.1	0.00968	0.0261	X	Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), transcript variant 1, mRNA.
ARID3A	-0.1	0.00969	0.0261	19	Homo sapiens AT rich interactive domain 3A (BRIGHT-like) (ARID3A), mRNA.
FAM83F	-0.21	0.00995	0.0267	22	Homo sapiens family with sequence similarity 83, member F (FAM83F), mRNA.
LPAR2	-0.1	0.01	0.0268	19	Homo sapiens lysophosphatidic acid receptor 2 (LPAR2), mRNA.
ABHD5	-0.1	0.0102	0.0271	3	Homo sapiens abhydrolase domain containing 5 (ABHD5), mRNA.
C5ORF28	-0.1	0.0105	0.0279	5	Homo sapiens chromosome 5 open reading frame 28 (C5orf28), mRNA.
PTPLAD2	-0.12	0.0106	0.0281	9	Homo sapiens protein tyrosine phosphatase-like A domain containing 2 (PTPLAD2), mRNA.
MYO1F	-0.1	0.0108	0.0284	19	Homo sapiens myosin IF (MYO1F), mRNA.
SEMA6B	-0.19	0.0109	0.0286	19	Homo sapiens sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6B (SEMA6B), mRNA.
MUC6	-0.25	0.0111	0.029	11	Homo sapiens mucin 6, oligomeric mucus/gel-forming (MUC6), mRNA.
CD19	-0.16	0.012	0.0308	16	Homo sapiens CD19 molecule (CD19), mRNA.

Continued on next page

Table A.1 – continued from previous page

Gene	logFC	p-value	q-value	CHR	Definition
WARS	-0.12	0.012	0.0308	14	Homo sapiens tryptophanyl-tRNA synthetase (WARS), transcript variant 1, mRNA.
SDHALP1	-0.1	0.012	0.0308		Homo sapiens succinate dehydrogenase complex, subunit A, flavoprotein pseudogene 1 (SDHALP1) on chromosome 3.
SLC25A23	-0.1	0.0121	0.0309	19	Homo sapiens solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 23 (SLC25A23), nuclear gene encoding mitochondrial protein, mRNA.
CXCR1	-0.12	0.0126	0.0318	2	Homo sapiens chemokine (C-X-C motif) receptor 1 (CXCR1), mRNA.
PLA2G2D	-0.11	0.0127	0.0321	1	Homo sapiens phospholipase A2, group IID (PLA2G2D), mRNA.
VPS41	-0.11	0.0129	0.0325	7	Homo sapiens vacuolar protein sorting 41 homolog (S. cerevisiae) (VPS41), transcript variant 1, mRNA.
NUBPL	-0.1	0.0131	0.0328	14	Homo sapiens nucleotide binding protein-like (NUBPL), mRNA.
ANKRD9	-0.17	0.0135	0.0334	14	Homo sapiens ankyrin repeat domain 9 (ANKRD9), mRNA.
ABCA7	-0.12	0.0137	0.0338	19	Homo sapiens ATP-binding cassette, sub-family A (ABC1), member 7 (ABCA7), mRNA.
ZNF223	-0.13	0.0139	0.0342		Homo sapiens zinc finger protein 223 (ZNF223), mRNA.
HBQ1	-0.16	0.0142	0.0348	16	Homo sapiens hemoglobin, theta 1 (HBQ1), mRNA.
SERPINA13	-0.33	0.0145	0.0353	14	Homo sapiens serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 13 (pseudogene) (SERPINA13), mRNA.
C9ORF130	-0.14	0.0145	0.0353		PREDICTED: Homo sapiens chromosome 9 open reading frame 130 (C9orf130), mRNA.
MATK	-0.11	0.0146	0.0353	19	Homo sapiens megakaryocyte-associated tyrosine kinase (MATK), transcript variant 3, mRNA.
MCOLN1	-0.15	0.0151	0.0365	19	Homo sapiens mucolipin 1 (MCOLN1), mRNA.
BCR	-0.23	0.0152	0.0366	22	Homo sapiens breakpoint cluster region (BCR), transcript variant 2, mRNA.
FAM116B	-0.15	0.0168	0.0396		Homo sapiens family with sequence similarity 116, member B (FAM116B), mRNA.
FOXO3	-0.11	0.0169	0.0399	6	Homo sapiens forkhead box O3 (FOXO3), transcript variant 2, mRNA.
GUK1	-0.12	0.0175	0.0409	1	Homo sapiens guanylate kinase 1 (GUK1), mRNA.
UCP2	-0.12	0.0177	0.0414	11	Homo sapiens uncoupling protein 2 (mitochondrial, proton carrier) (UCP2), nuclear gene encoding mitochondrial protein, mRNA.
EVI5	-0.11	0.0185	0.0426	1	Homo sapiens ecotropic viral integration site 5 (EVI5), mRNA.
REPS2	-0.12	0.0188	0.0432	X	Homo sapiens RALBP1 associated Eps domain containing 2 (REPS2), transcript variant 1, mRNA.
C15ORF39	-0.1	0.0198	0.0451	15	Homo sapiens chromosome 15 open reading frame 39 (C15orf39), mRNA.
ASCC2	-0.13	0.0203	0.046	22	Homo sapiens activating signal cointegrator 1 complex subunit 2 (ASCC2), mRNA.
CXCR5	-0.13	0.0203	0.046	11	Homo sapiens chemokine (C-X-C motif) receptor 5 (CXCR5), transcript variant 2, mRNA.
SLC25A37	-0.12	0.0204	0.0462	8	Homo sapiens solute carrier family 25, member 37 (SLC25A37), nuclear gene encoding mitochondrial protein, mRNA.
GFOD1	-0.1	0.0204	0.0462	6	Homo sapiens glucose-fructose oxidoreductase domain containing 1 (GFOD1), mRNA.
ICA1	-0.1	0.0219	0.0487	7	Homo sapiens islet cell autoantigen 1, 69kDa (ICA1), transcript variant 3, mRNA.
DNAJB2	-0.11	0.0219	0.0488	2	Homo sapiens DnaJ (Hsp40) homolog, subfamily B, member 2 (DNAJB2), transcript variant 2, mRNA.
TYW1	-0.12	0.0225	0.0499	7	Homo sapiens tRNA-yW synthesizing protein 1 homolog (S. cerevisiae) (TYW1), mRNA.

Full List of Enrichment Results for differential expression

Table A.2 Enriched Pathways for Differentially Expressed Probes (Complete)

Enriched Category	Library	Overlap	p-value	q-value
Up-regulated				
SRP-dependent cotranslational protein targeting to membrane (GO:0006614)	GO BP	43	0.0001	0.0001
viral transcription (GO:0019083)	GO BP	36	0.0001	0.0001
cotranslational protein targeting to membrane (GO:0006613)	GO BP	43	0.0001	0.0001
protein targeting to membrane (GO:0006612)	GO BP	48	0.0001	0.0001
establishment of protein localization to endoplasmic reticulum (GO:0072599)	GO BP	44	0.0001	0.0001
protein targeting to ER (GO:0045047)	GO BP	43	0.0001	0.0001
translational termination (GO:0006415)	GO BP	36	0.0001	0.0003
protein localization to endoplasmic reticulum (GO:0070972)	GO BP	43	0.0001	0.0004
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)	GO BP	41	0.0001	0.0006
translational elongation (GO:0006414)	GO BP	37	0.0001	0.0016
macromolecular complex disassembly (GO:0032984)	GO BP	41	0.0001	0.0026
cellular protein complex disassembly (GO:0043624)	GO BP	36	0.0001	0.0035
protein complex disassembly (GO:0043241)	GO BP	39	0.0001	0.0081
mRNA catabolic process (GO:0006402)	GO BP	48	0.0001	0.0084
protein localization to membrane (GO:0072657)	GO BP	54	0.0001	0.0093
establishment of protein localization to membrane (GO:0090150)	GO BP	53	0.0001	0.0105
viral life cycle (GO:0019058)	GO BP	38	0.0001	0.0121
nuclear-transcribed mRNA catabolic process (GO:0000956)	GO BP	47	0.0001	0.0123
translational initiation (GO:0006413)	GO BP	41	0.0001	0.0301
RNA catabolic process (GO:0006401)	GO BP	50	0.0001	0.0308
protein targeting (GO:0006605)	GO BP	53	0.0001	0.0426
translation (GO:0006412)	GO BP	56	0.0001	0.0559
protein localization to organelle (GO:0033365)	GO BP	62	0.0001	0.1022
cellular component disassembly (GO:0022411)	GO BP	50	0.0001	0.1185
establishment of protein localization to organelle (GO:0072594)	GO BP	56	0.0001	0.1252
single-organism cellular localization (GO:1902580)	GO BP	70	0.0002	0.3246
ribonucleoprotein complex biogenesis (GO:0022613)	GO BP	15	0.0005	0.9245
defense response to Gram-positive bacterium (GO:0050830)	GO BP	11	0.0011	1
cellular component biogenesis (GO:0044085)	GO BP	17	0.0011	1
hydrogen transport (GO:0006818)	GO BP	20	0.004	1
ribosomal subunit (GO:0044391)	GO CC	42	0.0001	0.0001
cytosolic large ribosomal subunit (GO:0022625)	GO CC	23	0.0001	0.0042
large ribosomal subunit (GO:0015934)	GO CC	25	0.0001	0.0052
ribosome (GO:0005840)	GO CC	40	0.0001	0.0139
cytosolic part (GO:0044445)	GO CC	43	0.0001	0.143
cytosolic small ribosomal subunit (GO:0022627)	GO CC	14	0.004	1
small ribosomal subunit (GO:0015935)	GO CC	17	0.0056	1
side of membrane (GO:0098552)	GO CC	24	0.0135	1
cell-cell junction (GO:0005911)	GO CC	23	0.0303	1
methyltransferase complex (GO:0034708)	GO CC	11	0.033	1
receptor complex (GO:0043235)	GO CC	18	0.0361	1
external side of plasma membrane (GO:0009897)	GO CC	17	0.0414	1
mitochondrial membrane part (GO:0044455)	GO CC	23	0.0471	1
histone deacetylase complex (GO:0000118)	GO CC	8	0.0482	1
structural constituent of ribosome (GO:0003735)	GO MF	42	0.0001	0.0013
calmodulin binding (GO:0005516)	GO MF	13	0.0042	1

Continued on next page

Table A.2 – continued from previous page

Enriched Category	Library	Overlap	p-value	q-value
small conjugating protein binding (GO:0032182)	GO MF	12	0.0065	1
monovalent inorganic cation transmembrane transporter activity (GO:0015077)	GO MF	20	0.0066	1
ubiquitin binding (GO:0043130)	GO MF	11	0.0139	1
hydrogen ion transmembrane transporter activity (GO:0015078)	GO MF	16	0.0174	1
inorganic cation transmembrane transporter activity (GO:0022890)	GO MF	23	0.0258	1
cysteine-type peptidase activity (GO:0008234)	GO MF	12	0.0463	1
cysteine-type endopeptidase activity (GO:0004197)	GO MF	8	0.0482	1
Ribosome Homo sapiens hsa03010	KEGG 2016	43	0.0001	0.0001
Parkinson's disease Homo sapiens hsa05012	KEGG 2016	24	0.0265	1
Oxidative phosphorylation Homo sapiens hsa00190	KEGG 2016	24	0.0359	1
Hematopoietic cell lineage Homo sapiens hsa04640	KEGG 2016	12	0.037	1
PGC SCZ	Pirooznia	24	0.0226	1
miR-137	Pirooznia	26	0.0277	1
BloodPlatelets customArray Gnatenko2	Blood	36	0.0231	1
turquoise M14 Nucleus HumanMeta	Brain	83	0.0001	0.0001
salmon M12 Ribosome HumanMeta	Brain	32	0.0001	0.0499
yellow M18 CTX	Brain	79	0.0001	0.1612
salmon M12 Ribosome MouseMeta	Brain	23	0.0046	1
green M10 GlutamatergicSynapticFunction CTX	Brain	44	0.0099	1
greenyellow M6 GlutatmatergicSynapse MouseMeta	Brain	25	0.0233	1
red M11 Neuron HumanMeta	Brain	34	0.0374	1
brown pyramidalNeurons Layer5/basolateralAmygdala Sugino/Winden	Brain	31	0.0476	1
Substantia Innominata localMarker(top200) IN Basal Forebrain	HBA	10	0.0484	1

Differential Expression of Medication Subgroups

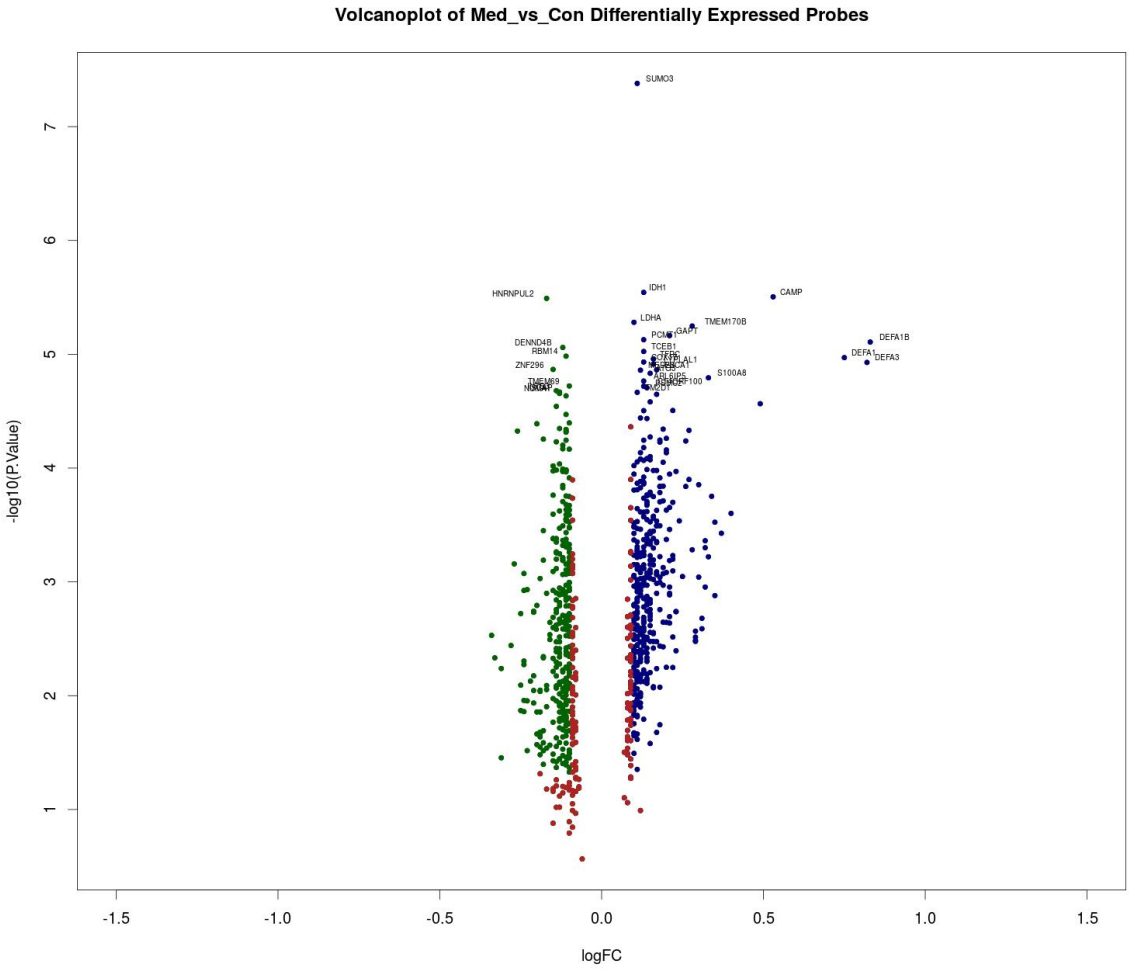


Figure A.1 Visualizations of Differential Expression (Medicated group)

Shows a volcanoplot of limma differential expression results between medicated FEP and healthy controls (HC) samples. Blue probes are up regulated in FEP and green probes are down-regulated. Red probes are considered unchanged either due to low q-value, or low differential expression.

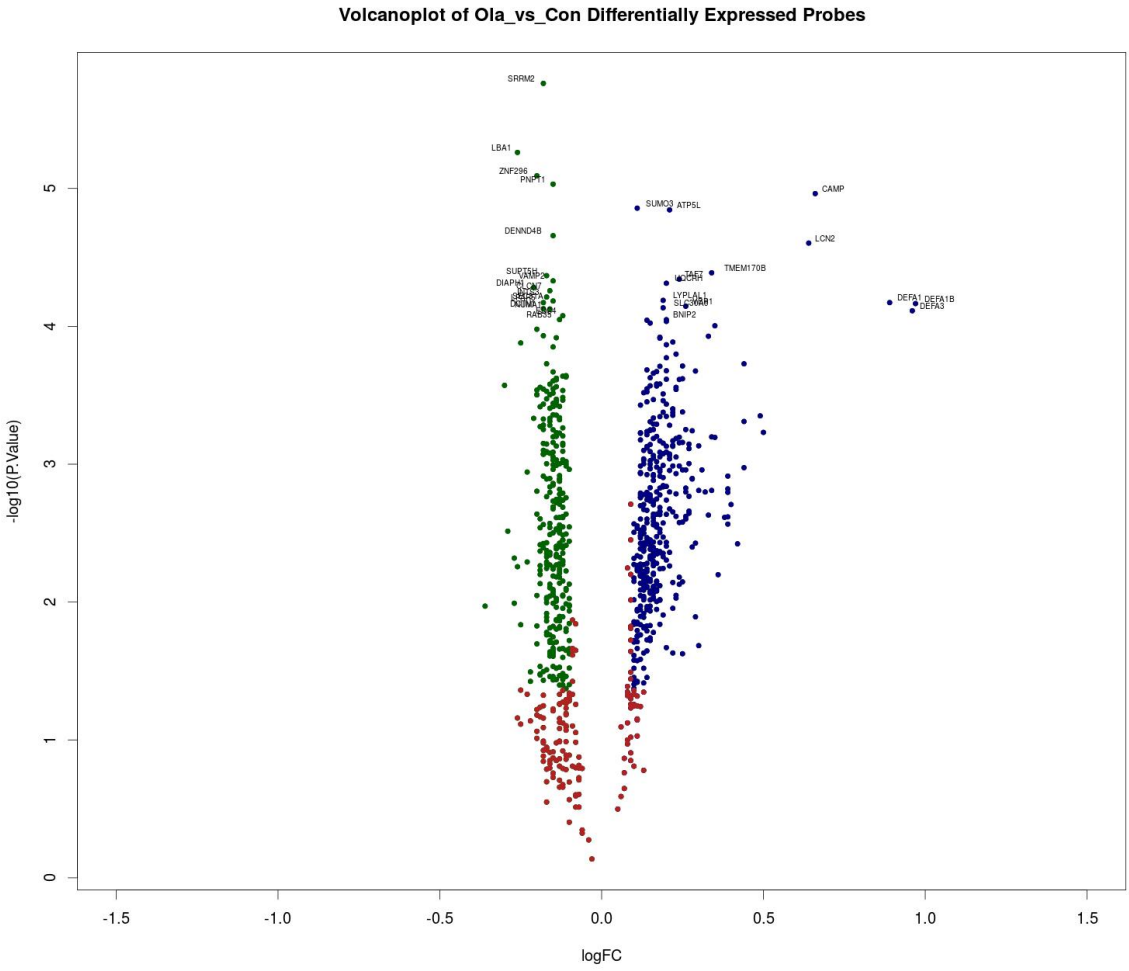


Figure A.2 Visualizations of Differential Expression (Olanzapine)

Shows a volcanoplot of limma differential expression results between Olanzapine medicated FEP and HC samples. Blue probes are up regulated in FEP and green probes are down-regulated. Red probes are considered unchanged either due to low q-value, or low differential expression.

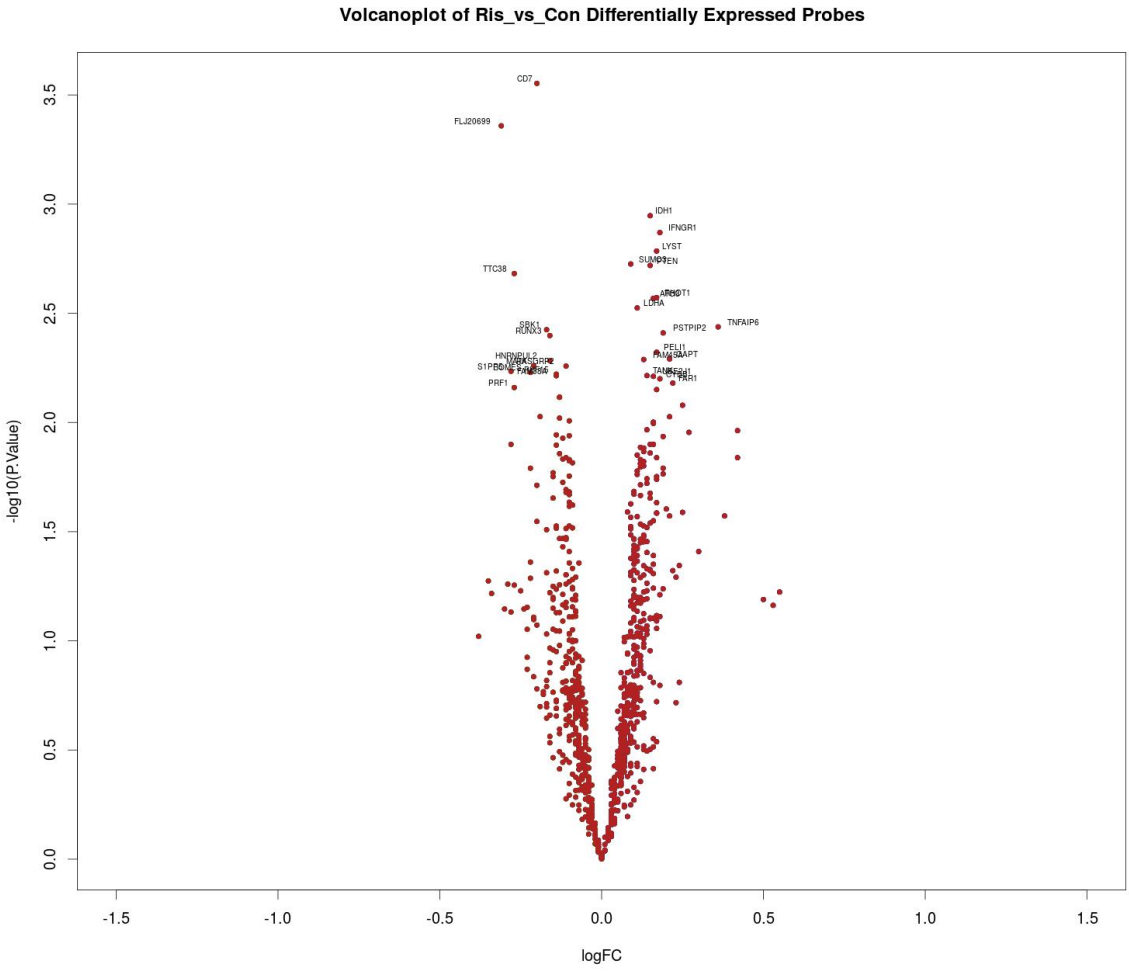


Figure A.3 Visualizations of Differential Expression (Risperidone)

Shows a volcanoplot of limma differential expression results between Risperidone medicated FEP and HC samples. Blue probes are up regulated in FEP and green probes are down-regulated. Red probes are considered unchanged either due to low q-value, or low differential expression.

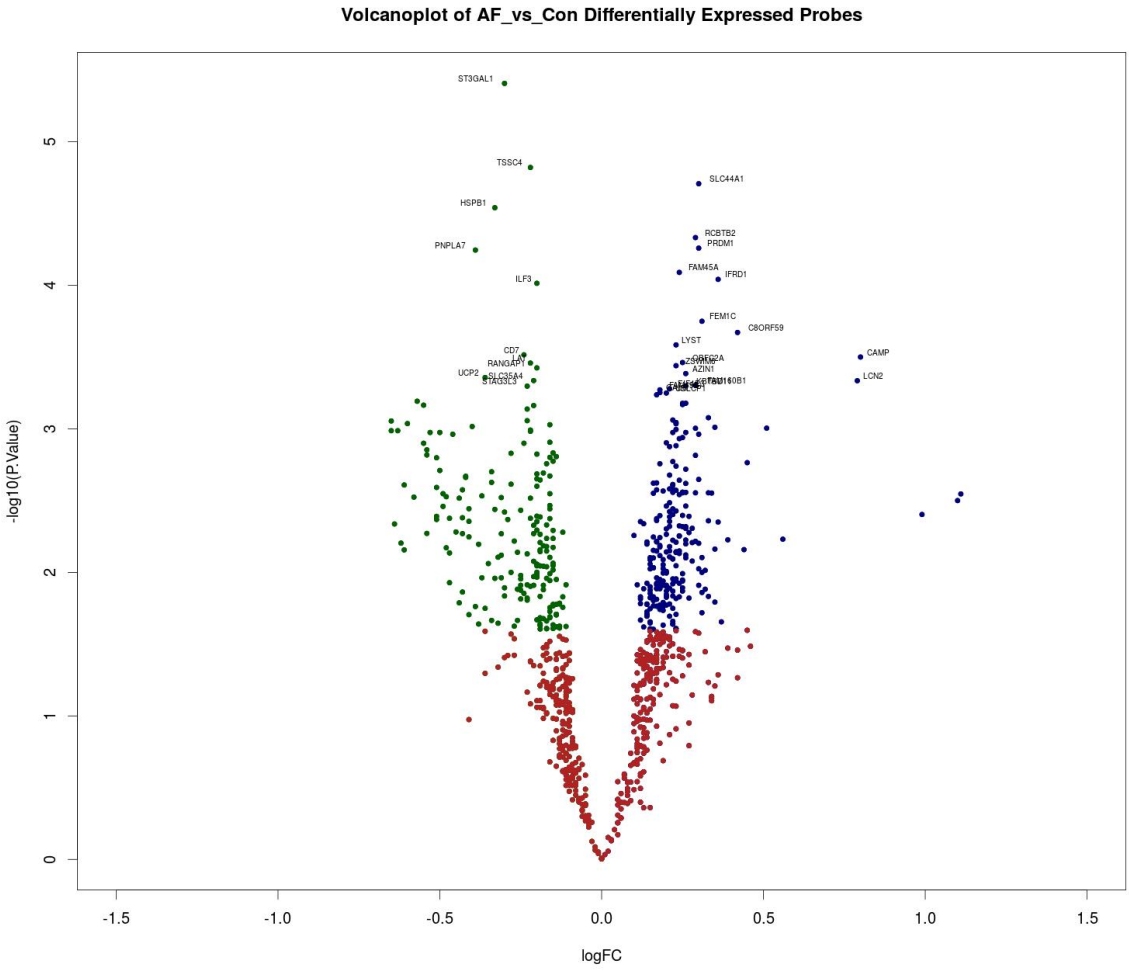


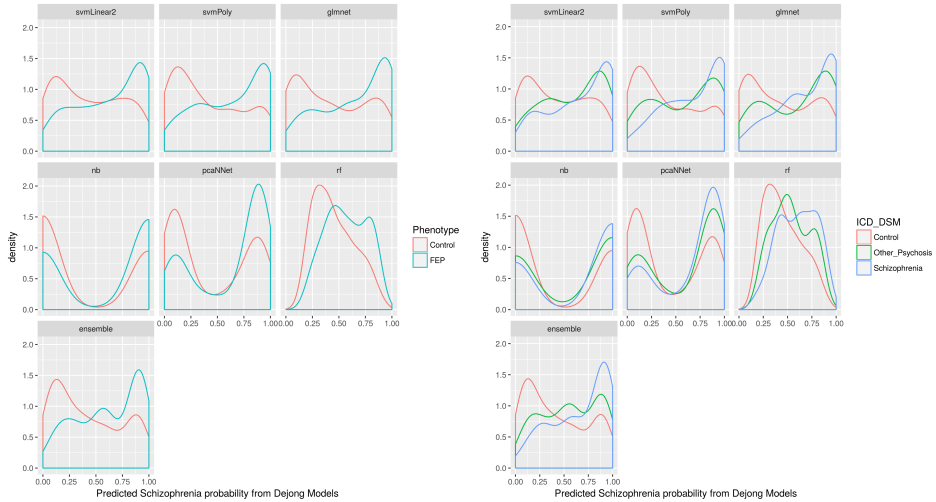
Figure A.4 Visualizations of Differential Expression (Anti-psychotic Free)

Shows a volcanoplot of limma differential expression results between medication free FEP and HC samples. Blue probes are up regulated in FEP and green probes are down-regulated. Red probes are considered unchanged either due to low q-value, or low differential expression.

Appendix B

Supplementary Material: Chapter 5

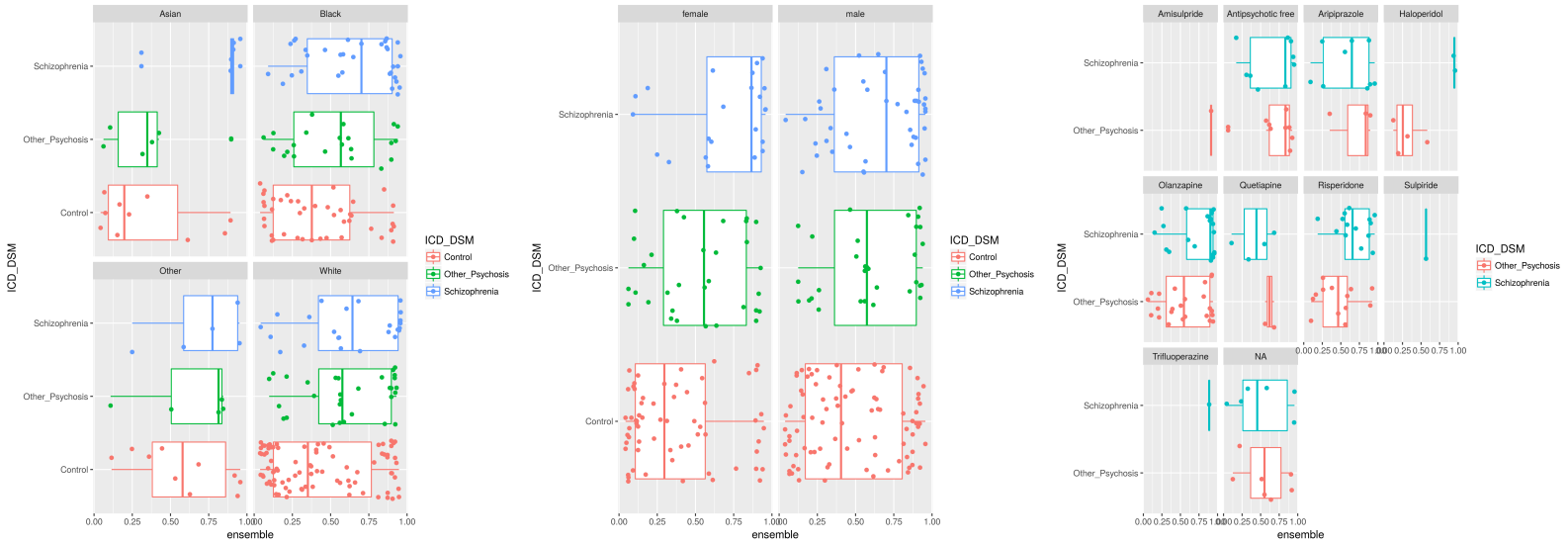
Additional Figures visualising predictive performance in GAP



(a) Stratified by Control and FEP (b) Stratified by Diagnosis

Figure B.1 Density plots for Glmnet Gene expression model.

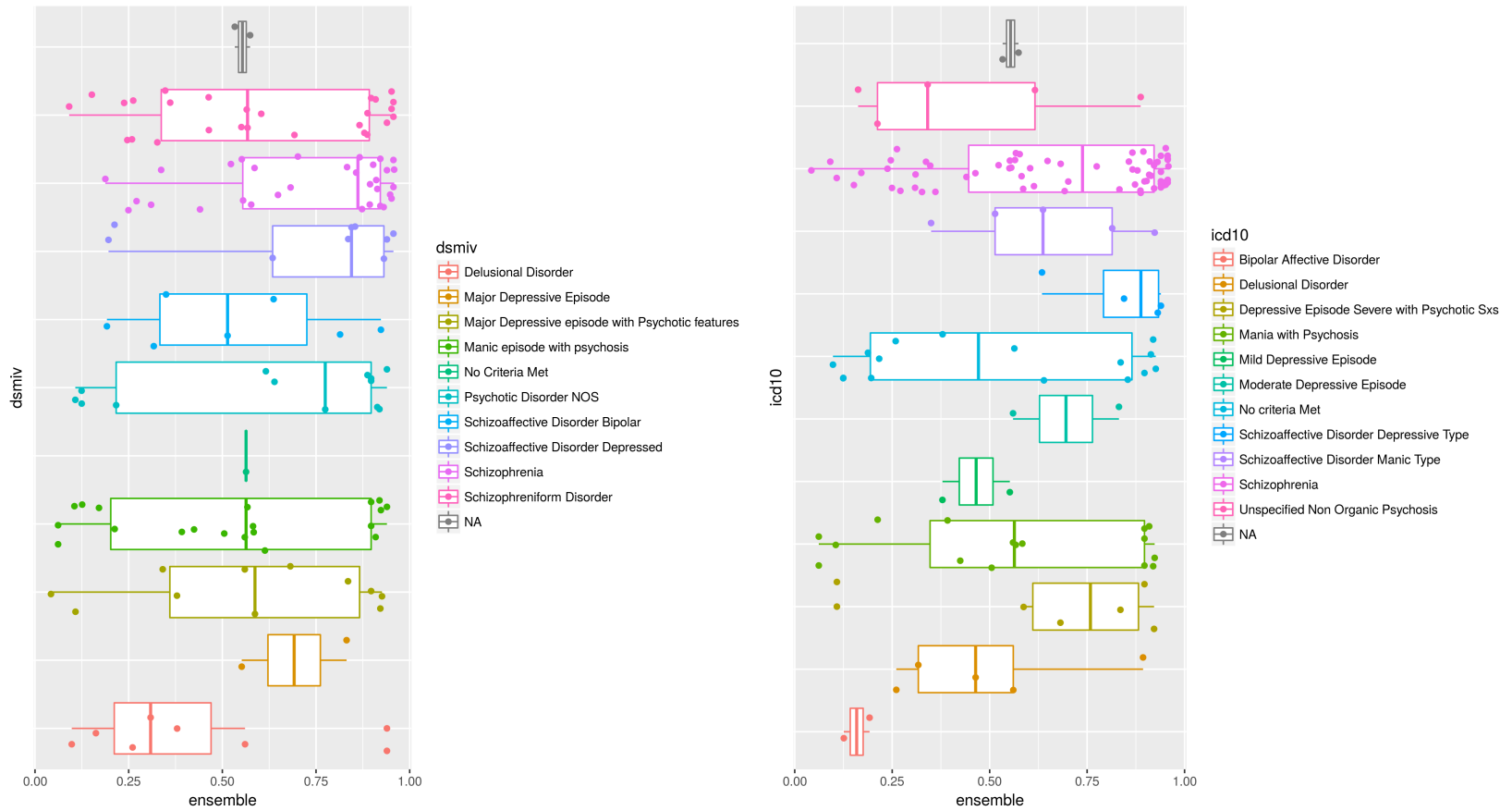
Density plots of Probability predictions for GAP samples as Controls based on all Dejong Models from first training fold. Values are between 0 and 1, with higher values representing a higher confidence that the sample belongs to a patient. (a) Shows data split by Control and First Episode Psychosis (FEP). (b) Shows the same data, split with first episode psychosis samples split across Schizophrenia samples, and Other Types of Psychosis.



(a) Ethnicity (b) Gender (c) Medication

Figure B.2 Additional Demographics (Chapter 5)

Boxplots visualising accuracy of Dejong ensemble model applied to GAP data, while incorporating Ethnicity, Gender or Medication. In each case samples are stratified by Diagnosis as described in Chapter 5. The x-axis indicates confidence of the ensemble model that a sample is a patient. A value above 0.5 means a samples is classified as a Schizophrenia patient, while a value below 0.5 indicates that a sample is classified as control. (a) Shows four plots based on Ethnicity. Schizophrenia samples are classified with more confidence as patients. (b) Shows two plots based on Gender. (c) Shows 10 plots based on medication. The y-axis variation for individual points within a group, is unrelated to the data, and is purely a visual aid to clarify the distribution of points across the x-axis.



(a) DSMIV

(b) icd10

Figure B.3 Predictive Probability split by all ICD-10 and DSM-IV diagnoses

Boxplots visualising Predictive accuracy of Dejong ensemble model applied to GAP data, with samples stratified by diagnosis based on ICD-10 or DSM-IV. The x-axis indicates confidence of the ensemble model that a sample is a patient. A value above 0.5 means a samples is classified as a Schizophrenia patient, while a value below 0.5 indicates that a sample is classified as control. (a) Shows data stratified using DSM-IV diagnostic categories. (b) Shows data stratified using ICD-10 diagnostic categories. The y-axis variation for individual points within a group, is unrelated to the data, and is purely a visual aid to clarify the distribution of points across the x-axis.

Tables of results including GAP model predictions in Dejong data

Table B.1 AUC for split 1 of GAP and Dejong Models tested on both datasets

Built on	Tested on	Machine Learning Model						
		svmLinear	svmPoly	glmnet	nb	pcaNNet	rf	ensemble
GAP	GAP	0.62	0.65	0.61	0.62	0.61	0.64	0.63
	Dejong	0.68	0.7	0.74	0.71	0.71	0.73	0.72
Dejong	Dejong	0.82	0.87	0.87	0.86	0.84	0.84	0.84
	GAP	0.67	0.69	0.69	0.66	0.68	0.66	0.68

Table of AUC for all 7 machine learning models in the first split of both Dejong and GAP Models. The first column shows the data set used to build models, the second column shows the dataset used for testing. Models built on GAP were created using the same methods as described for the Dejong models. Data for models built on Dejong is identical to what is shown in Chapter 5. Results show that models built on GAP, in all cases, are more accurate at predicting the external Dejong data, than the internal GAP validation data. In addition GAP data is predicted more accuratly by models trained on Dejong Data, than by models trained on GAP.

Table B.2 Analysis of Classification predictions in Dejong data from GAP models

GAP_Data	Method	Accuracy	Balanced.Accuracy	Sensitivity	Specificity
Split_1	svmLinear2	0.614	0.622	0.453	0.792
	svmPoly	0.658	0.66	0.623	0.698
	glmnet	0.649	0.649	0.632	0.667
	nb	0.639	0.638	0.642	0.635
	pcaNNet	0.624	0.625	0.594	0.656
	rf	0.649	0.648	0.66	0.635
	ensemble	0.644	0.651	0.509	0.792

Table showing performance of models trained on GAP data, in the Dejong dataset. These results mirror Table 5.5, where data is presented on performance of Dejong models in GAP data.